ASSESSMENT OF GROUP PREFERENCES AND GROUP UNCERTAINTY FOR DECISION MAKING

SOCIAL SCIENCE RESEARCH INSTITUTE
UNIVERSITY OF SOUTHERN CALIFORNIA

David A. Seaver

ADVANCED DECISION TECHNOLOGY PROGRAM

CYBERNETICS TECHNOLOGY OFFICE
DEFENSE ADVANCED RESEARCH PROJECTS AGENCY
Office of Naval Research • Engineering Psychology Programs

The objective of the Advanced Decision Technology Program is to develop and transfer to users in the Department of Defense advanced management technologies for decision making. These technologies are based upon research in the areas of decision analysis, the behavioral sciences and interactive computer graphics. The program is sponsored by the Cybernetics Technology Office of the Defense Advanced Research Projects Agency and technical progress is monitored by the Office of Naval Research – Engineering Psychology Programs. Participants in the program are:

Decisions and Designs, Incorporated
The Oregon Research Institute
Perceptronics, Incorporated
Stanford University
The University of Southern California

Inquiries and comments with regard to this report should be addressed to:

Dr. Martin A. Tolcott
Director, Engineering Psychology Programs
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

or

LT COL Roy M. Gulick, USMC
Cybernetics Technology Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

(9) TECHNICAL REPORT SSRI-76-4
Jul 75 - Sep 76

(6) # ASSESSMENT OF GROUP PREFERENCES AND GROUP UNCERTAINTY FOR DECISION MAKING.

(10) by
David A. Seaver

(15) Sponsored by
Defense Advanced Research Projects Agency
N00014-76-C-0074 ARPA Order No. 3052
Under Subcontract from
Decisions and Designs, Incorporated

(11) June 1976

(12) 65 p.

SOCIAL SCIENCE RESEARCH INSTITUTE
University of Southern California
Los Angeles, California 90007

SUMMARY

Decision analysis has rapidly become an accepted tool for aiding decision makers to make optimal decisions. The use of decision analysis involves the quantification of the decision maker's preferences and opinions as utilities and subjective probabilities respectively. However, the formal theory underlying the development of decision analysis is based on the decision maker's being a single identifiable individual. Often groups rather than individuals serve as decision makers. Even when a single individual functions as the decision maker, a group may be called upon to provide the inputs necessary for making decisions. In these situations, group utilities and probabilities must be determined. The obvious approach to determining group utilities and probabilities is somehow to combine the judgments of the individuals in the group into a group judgment. Theoretical research, however, has proved that no really satisfactory method for combining individual utilities or probabilities into a group utility or probability exists. The purpose of this report is to explore the possibilities that exist for determining group utilities and probabilities, focusing on the advantages and disadvantages of the various procedures.

The report begins by assessing the current state of the art with respect to determining group preferences and utilities. Three specific possible methods for combining individual preference or utility functions into group preference or utility functions are explored. All suffer from rather severe disadvantages such as restrictive applicability or violation of Pareto optimality. Certain experimental conditions that may reduce disagreement and, therefore, lead to a greater chance of unanimity among group members are also discussed.

There are two general procedures for forming group probability judgments: mathematical aggregation procedures and behavioral methods. The mathematical aggregation procedures depend on a mathematical formula for determining the group probabilities from the individual probabilities. Several possibilities exist, but those with the best underlying theory typically cannot be used in practical situations because of the difficulty in determining some of the necessary inputs.

The behavioral methods utilize interaction or communication among the group members to try to reduce the disagreement among group members so a consensus will result. The most widely used methods depend on highly structured communication to allow the group to profit from certain advantages of group interaction that are well-documented by social psychological research.

Since none of the procedures reviewed for forming group utilities or probabilities is completely acceptable on a theoretical level, choice among any set of applicable procedures should be based on empirical observations of the quality of the resulting group judgments. However, since very little empirical research has been done in this area, few conclusions about the relative effectiveness of the different methods can be drawn.

CONTENTS

FIGURES

ACKNOWLEDGMENTS

Support for this research performed by Social Science Research Institute was provided by the Advanced Research Projects Agency of the Department of Defense and was monitored under Contract N00014-76-C-0074 with the Office of Naval Research, under subcontract from Decisions and Designs, Inc.

The author would like to thank Dr. Ward Edwards for many suggestions leading to the improvement of this paper.

I. INTRODUCTION

Out of a diverse background has grown a set of theories, collectively called decision theory, that describe how people do and should make decisions. For the most part this development has been concerned with individual decision making (see Rapoport and Wallsten, 1972, for a recent review). Yet decisions are often made by groups. Even when a single person can be designated as the decision maker, groups are often relied upon for advice that serves as a direct input into the decision making process.

One widely accepted theory and applied technology, decision analysis (cf. Raiffa, 1968), prescribes how a decision should be made in a situation simply characterized by a set of alternative courses of action, a mutually exclusive and exhaustive set of events or states of the world, and a set of outcomes which accrue for each alternative depending on which event occurs. In this situation the choice among the alternatives should depend upon the decision maker's preferences for the outcomes and opinions about which event will occur. Decision analysis provides a set of tools for quantifying the preferences as utilities, the opinions as probabilities, and a decision rule--maximize expected utility--for choosing among the alternatives on the basis of the quantified preferences and opinions. Although the mathematical development is elegant and sophisticated, the result is rather straightforward to apply in many situations with a single identifiable decision maker.

But what happens when a group is vested with the decision making responsibility? The board of directors may disagree about the weight that should be given to public image in developing an overall corporation utility function. A parole board may disagree about the probability that a prisoner being considered for parole will commit another crime. Decision making groups are pervasive in both the private and public sectors of our society. Thus, as decision analysis becomes a more accepted tool for aiding decision makers, there is a greater need for the development of normative theories of group decision making and the technology for their application. Can diverse individual preferences and opinions be combined to yield utilities and probabilities that represent the individuals? Is the same decision rule applicable for groups that has come to be accepted as rational for individuals?

An obvious approach to this problem is to look for procedures for aggregating the individual judgments into a group judgment, for example, averaging the individual utilities and probabilities. This approach has led to unsatisfactory theoretical results. Arrow (1951) proved the famous impossibility theorem that states there is no aggregation rule for combining individual preferences into a group preference that satisfies a reasonable set of requirements. Dalkey (1972) has proved a similar result for probabilities. Furthermore, a problem with the decision rule arises in some situations where there may be a conflict between maximizing expected utility and satisfying the Pareto optimality condition*, another widely ac-

* The Pareto optimality condition states that an alternative is Pareto optimal if there is no other alternative that is at least as good for

$P_1(\theta_i)$ = .6 .4

	θ_1	θ_2	EU_1
a_1	1.0	0.0	.60
a_2	.4	.8	.56

Individual 1

$P_2(\theta_i)$ = .3 .7

	θ_1	θ_2	EU_2
a_1	0.0	1.0	.70
a_2	.2	.6	.48

Individual 2

$P_G(\theta_i)$ = .45 .55

	θ_1	θ_2	EU_G
a_1	.5	.5	.50
a_2	.3	.7	.52

Group

Figure 1: Violation of Pareto Optimality when Individual Probabilities and Utilities are Averaged

cepted normative condition for group choice. Raiffa (1968, Ch. 8) and Dalkey (1975) give examples similar to the following. Figure 1 illustrates a group of two individuals choosing between the two alternatives, a_1 and a_2. The outcomes received by the individuals depend upon which of the two states of the world, θ_1 or θ_2, occurs. The decision matrix for each individual is given (entries in the matrices are utilities) along with the subjective probabilities for the states of the world. Clearly, each individual should favor alternative a_1. If the individual probabilities and utilities are averaged to arrive at group probabilities and utilities, the given matrix becomes the decision matrix for the group. Under these conditions the group should choose alternative a_2. Thus the maximization of expected utility by the group conflicts with choosing the Pareto optimal alternative, a_1.

Such paradoxes are not simply due to the aggregation procedure, in this case averaging. Raiffa (1968) notes a theorem without proof by Richard Zeckhauser which states:

> "Suppose you announce a group procedure that (1) combines utility and probability functions separately, and (2) does not single out one individual to dictate the group utility and probability assignments. Then you can always concoct an example such that each of your experts will agree on which act to choose but where your group procedure will lead you to a different conclusion" (p. 230).

Bacharach (1975) has recently proved a similar theorem.

The circumstances are equally complex when a single decision maker seeks expert advice. Here an additional question arises, namely, will a group of experts provide better advice than a single expert? If the decision maker decides to use several experts, then the above-mentioned problem arises. These negative results cannot simply be accepted and the problems dismissed as having no solutions. Decisions are and will continue to be made by groups. Our goal should be to find ways to aid and improve the decision making process. The value of a technology is measured not by answering: Is it correct? but rather by answering: Is it better than any available alternative?

This approach will be taken in this paper, which reviews the theoretical and experimental literature on groups making quantitative judgments of the type needed in decision analysis. The major emphasis of this review

---

*every member of the group and "better" for at least one member. Only Pareto optimal alternatives should be considered for choice by the group. In the context of the present discussion, "better" means higher expected utility. A more restrictive definition would define "better" as higher quality for each outcome determined by the possible events. Such a restrictive definition reduces the usefulness of the Pareto optimality condition for group choice.

will be on probability and related types of judgments where uncertainty is involved. A brief discussion of decision analytic work on the group utility problem will also be included for the sake of completeness. Consideration of group probability assessment will focus on two subtopics: (1) individual versus group judgments, and (2) the processes by which the group judgment can be reached. The second topic is, of course, included in the first since the group judgments must be formed in some manner to enable comparison with individual judgments. However, it seems desirable to keep the two questions conceptually distinct.

Certain related topics have been specifically excluded from this paper including discussions of particular techniques for assessing individual probabilities and utilities. Excellent discussions of the techniques for assessing subjective probabilities and utilities may be found in Spetzler and Stael von Holstein (1975) and von Winterfeldt and Fischer (1975) respectively.

The voluminous work in welfare economics, which is directly related to the group utility problem, has not been included. Much of this research stems from Arrow's (1951) theorem, including reformulations, weakening of the assumptions, and use of cardinal utility. Books by Arrow (1963), Fishburn (1973) and Pattanaik (1971), along with a chapter by Luce and Raiffa (1957, Ch. 14) will serve as useful guides to this topic.

Game theory has also been omitted from this review. Rapoport provides a complete discussion of two-person (1966) and n-person (1970) game theory. Luce and Raiffa (1957) is also an excellent source.

Finally, only the surface of related research in social psychology has been skimmed. The risky shift has received considerable experimental emphasis (see Clark, 1971 and Vinokur, 1971 for reviews), but has not been included here due to the bias of this review toward normative rather than descriptive theory. The subject of group dynamics and processes has been explored only to the extent necessary to help explain the results of different approaches to group judgments.

II. ASSESSING GROUP PREFERENCES AND UTILITIES

This section will concentrate on recent attempts to develop formal procedures for aggregating individual preferences or utilities into a group preference or utility function. As defined here, the distinction between preferences and utilities is mathematically quite simple: preferences are measured on an ordinal scale while utilities are measured on a cardinal scale, that is, a scale of at least interval length. A utility function is, therefore, a special case of a preference function. Theoretically, utilities are needed for calculating expected utilities in choice situations where the outcomes are known only probabilistically. If the outcomes of decisions are known with certainty, only preferences are needed for decision making. Thus, for most decisions the subjective values must be measured as utilities, since decisions where the outcomes are known with certainty are rare. However, in practice this distinction is of less importance because preferences measured on an ordinal scale can be transformed into utilities measured on a cardinal scale via lottery procedures that define an origin and unit of measure for the utility function and take into account the attitude toward risk (von Winterfeldt and Fischer, 1975).

Since much of the work reported in this section uses Arrow's (1951) theorem as a point of departure, I will first state the five conditions that he proved to be inconsistent, that is, there is no rule for aggregating individual preference orderings into a group ordering that satisfies all of these conditions. Notice that Arrow deals only with ordinal preferences, not cardinal utilities.

Arrow's conditions are:

1. There are at least two individuals in the group, at least three alternatives to choose among, and a complete group ordering exists for all possible profiles of individual orderings.

2. If, for a certain profile of individual orderings, the group ordering asserts x is preferred to y, then x must still be preferred to y in the group ordering if all individual orderings of alternatives other than x remain the same and each individual's ordering of x with respect to any other alternative remains the same or is modified in favor of x.

3. If a profile of individual orderings is modified in such a manner that each individual's ordering among a subset of the alternatives remains the same, the group ordering for the original and modified individual profiles must be identical for that subset of alternatives.

4. For each pair of alternatives x and y, there is some profile of individual orderings such that the group prefers x to y.

5. There is no individual with the property that whenever he prefers x to y, the group also prefers x to y regardless of the other individuals' orderings.

Arrow's theorem does not necessarily mean that in a specific social choice situation with a given set of alternatives to be considered and a given profile of individual orderings, there is no procedure for aggregating the individual orderings into a group ordering. It only says that there is no such procedure which will work in all situations, for all profiles of individual orderings. Thus, in practice, a check should first be made to see if in the specific situation under consideration, there may be some procedure, for example, majority rule, that will yield a complete social ordering.

Of Arrow's conditions, only condition 3 has been subjected to any substantial amount of criticism (Luce and Raiffa, 1957; Pattanaik, 1971). This is the condition that is violated by most commonly used aggregation rules, for example, the sum of ranks. Although some counterarguments to condition 3 do exist, it is usually accepted as a seemingly reasonable requirement.

II. A. Restricted Individual Preference Orderings

A possibility that has been investigated for thwarting the impact of Arrow's conclusion is to weaken condition 1. This is not a result of criticisms of the reasonableness of the condition, but because it may simply be a stronger condition than is necessary in many situations. Two possibilities have been explored: relaxing the requirement for a complete group ordering, and restricting the set of profiles of individual orderings for which the group ordering is defined.

Pattanaik (1971) discussed the difference between social decision functions and social welfare functions, the latter being what Arrow used in his theorem. Social decision functions do not require a complete social ordering, only that some subset of the original alternatives can be chosen that is in some sense optimal. Such a formulation does not eliminate the problem since there may be more than one alternative in the chosen set. In any case, Arrow's theorem applies to social decision functions as well as to social welfare functions.

Another proposal for weakening the requirements for the group references is to drop the transitivity requirement. At first thought this seems entirely irrational since transitivity is the cornerstone of individual choice theories. Fishburn (1970) provided an argument and some intuitive examples showing why transitivity may not be desirable. Essentially, he argues that in many situations a choice must be made between satisfying social transitivity and using a majority rule decision function; and at least to some people, dropping the transitivity requirement is more acceptable.

The majority rule decision function plays a major role in investigations of the type of restrictions on the profiles of individual orderings that will lead to social decision functions or social welfare functions. It is often used in practical situations and is well-established in democratic societies. In addition, it satisfies all of Arrow's conditions except condition 1. Fishburn (1973) and Pattanaik (1971), among others, extensively discuss the restrictions under which majority rule will lead to a social decision function or a social welfare function.

One restriction, called Value Restriction (VR) by Pattanaik (1971), is a generalization of the single-peaked conditions of Black (1948) and Coombs (1954). VR holds if, for a triple of alternatives among which all individuals are not indifferent, there is one alternative such that for all individuals, it is not given the worst value, or it is not given the best value, or it is not given the medium value. Although this condition seems to be quite confusing, consideration of what triples violate VR may be elucidating. A triple violates VR if for each of the alternatives in the triple, there is at least one individual who gives the alternative the best value, one who gives it the medium value, and one who gives it the worst value. For example, if alternatives a, b, and c are being considered by three individuals with the following preference orders, abc, bca, and cab, VR is violated. Each of the alternatives is best, medium, and worst in one of the preference orderings.

Using the VR condition, Pattanaik proved first, that the majority rule leads to a social decision function, that is, the choice set is nonempty, if VR holds for every triple in the Pareto-optimal subset of alternatives. Pattanaik proved second, that the majority rule will yield a social welfare function, that is, a complete ordering of alternatives, if VR is satisfied for every triple of alternatives and there is an odd number of individuals who are not indifferent to all alternatives in the triple.

It seems that VR can reasonably be expected to be satisfied in certain situations. Even if it is not satisfied for all triples by all individuals, practical indications suggest that majority rule will still yield a useful social ordering of the alternatives particularly if the number of individuals is large. In particular, if there is a single attribute underlying the preferences of individuals, this restriction can be expected to be satisfied. For example, in choosing among presidential candidates, if all individuals' preferences were dominated by the liberal-conservative attribute of the candidates, VR most likely would be satisfied for most triples of candidates. With the triple consisting of a liberal (x), a moderate (y), and a conservative (z) candidate, there are thirteen possible individual preference orders, including indifference among all candidates. Using P to indicate strict preference and I to indicate indifference, the possible orderings are:

1. xPyPz	6. zPyPx	11. zPxIy
2. xPzPy	7. xPyIz	12. zIxPy
3. yPxPz	8. xIyPz	13. xIyIz
4. yPzPx	9. yPxIz	
5. zPxPy	10. yIzPx	

Figures 2 (a), (b), and (c) show which of these orderings satisfy the not-worst (single-peaked), not-best (single-caved) and not-medium, conditions of VR, respectively, for alternative y.

The assumption of an odd number of individuals, which is necessary to insure a complete social ordering, is, in practice, not very restrictive. It is used to guarantee that there are no ties under the majority rule decision function. If the number of individuals who are not indifferent to all alternatives in the triple is large, there will be little chance of a tie even if the number of individuals is even.

II. B. Anchored Preference Scales

Dalkey (1975) suggested that the paradox of Arrow is the result of an overstrict notion of ordinal scales. He argued that the role of reference objects for ordinal scales has been overlooked in dealing with social welfare functions. This role is illustrated by the Mohs hardness scale, where the ordering relation is "scratches." Typically, there is a fixed reference set so the hardness of any stone can be measured by which stones it scratches and which stones scratch it. If the fifth hardest reference object scratches the stone being measured, and that stone in turn scratches the fourth hardest stone in the reference set, the hardness of the stone is said to be between 4 and 5. These numbers, however, are purely ordinal. Dalkey calls such a scale with the ordinal relation R and a fixed set of reference objects an anchored scale. The scale value S(x) of any object, x, being measured is defined to be the rank order of the highest reference object, a, such that xRa.

The group preference problem is then formulated in terms of anchored scales as follows: Each individual has a fixed anchor set (not necessarily the same for all individuals). The group anchor set is the cartesian product of all individual anchor sets. Each individual's scale, S_j, is derived from his preference relation, R_j. The group scale is then formed by some function, $F(S_1, \cdots, S_n)$, and the group preference order is determined by the group scale. Care must be taken in distinguishing between preference orderings and scales. The individual's preference orderings determine the individual scales, but the group scale determines the group preference ordering.

With Arrow's conditions then expressed in terms of the scale values rather than preferences, and the objects in the anchor set assumed to be exempt from these conditions, Dalkey demonstrated that a group preference scale does exist which satisfies these conditions. To do this, he used the

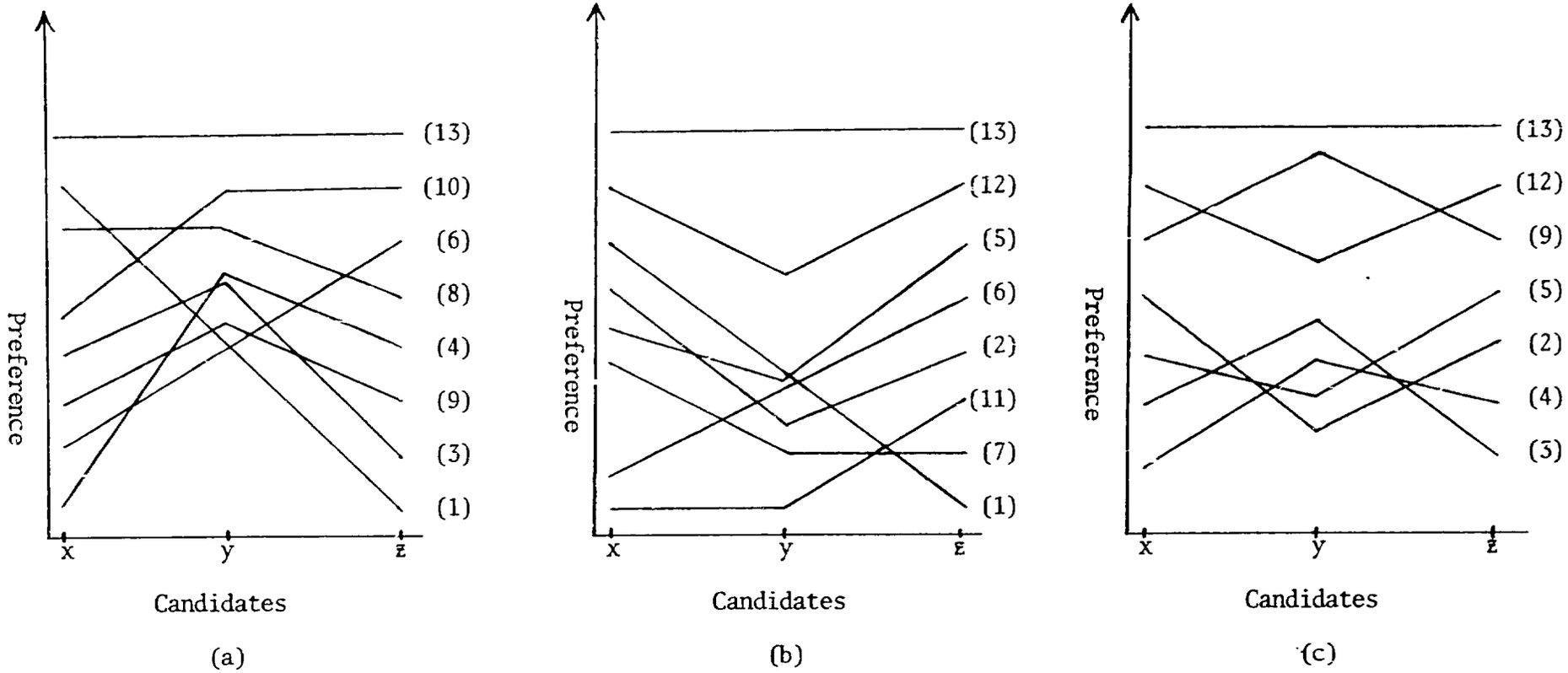

Figure 2: Examples of Not-Worst (a), Not-Best (b), and Not-Medium (c) VR.

sum of the individual scales as the group scale. This procedure assigns a numeric value to each alternative, so since the arithmetic inequality is a complete order, this yields a complete social preference order satisfying condition 1. Condition 2 is obviously satisfied by the summation procedure. The group scale value does not change when only subsets of the alternatives are considered, so Condition 3 is satisfied. Condition 4 is satisfied, since, by Condition 1, there is some x and y such that $S_i(x) \geq X_i(y)$ for all i, so $\Sigma S_i(y)$. To show that Condition 5 is satisfied consider that condition 1 requires that there are at least three potential rank order numbers. Then there is a pair of alternatives, x and y, such that for one individual k, $(S_k(s) = S_k(y) + 1$ and for all individuals $j \neq k$, $S_j(y) = S_j(x) + 2$. The group scale value of y is $S(y) = \Sigma S_i(y) = \Sigma S_i(x) + 2(n-1) - 1$, so the group prefers y to x while individual k prefers x to y. Since the group preference scale exists and defines a complete ordering, this ordering can be taken as the group preference ordering.

The crucial point in Dalkey's formulation of the group preference problem that allows him to find an aggregation procedure satisfying the five conditions is that the anchor set does not need to satisfy the conditions. Clearly, this allows for what would otherwise be a violation of the condition of independence from irrelevant alternatives. Preferences expressed in terms of scale values are not independent of the anchor set.

The biggest problem in using this procedure to obtain a group preference ordering is the choice of the anchor sets. The resulting ordering may be very sensitive to this choice. Since the anchor set does not have to satisfy the five conditions, it would seem to be desirable not to include alternatives that are actually under consideration in the anchor set. Additional problems arise since different individuals can have different anchor sets. In fact, nothing requires individuals to have the same anchor sets or even the same number of reference objects in the anchor sets. Thus, if one individual has only two reference objects in the anchor set and another individual's anchor set includes fifty reference objects, the preferences of the second individual will obviously swamp the preferences of the first individual. Note also that any individual can possibly assign the same scale value to two alternatives between which he or she has a strict preference. The scale values certainly are not guaranteed to be as sensitive as a complete preference ordering. In fact, by using the sum-of-anchored-scales procedure to produce the group scale, two alternatives may be judged indifferent by the group when one is strictly preferred to the other by each member of the group, a violation of Pareto optimality.

II. C. Cardinal Utility and Interpersonal Comparisons

A point crucial to decision analysis that is not satisfied by any of the procedures previously discussed in this section is that in many decision situations ordinal scales of preference are not strong enough for decision making. For the expected utility maximization decision rule to be used in situations where the outcomes resulting from a decision are known only probabilistically, the utility functions must be of at least interval strength. If the ordinal preferences are to be used for calculating expected utilities, they must first be transformed onto an interval scale. Since such a transformation is guaranteed to exist (procedures for making the transformation are discussed in von Winterfeldt and Fischer, 1975), the ordinal preference scales may still be useful. However, problems may arise in making the transformation from ordinal to interval scales. The transformation depends on the risk attitude of the decision maker whose utilities are being represented. If the decision maker is a group, there is every reason to believe that different members of the group will have different risk attitudes. Thus, it must be determined what risk attitude represents the entire group. Even if all members have the same risk attitude, it can be argued that the group as a whole should be less risk averse than each of the individual members, since they now share the risk. Raiffa (1968, pp.188-220) provides an elucidating discussion and formal analysis of risk sharing, particularly in the context of deriving group utility functions.

To avoid the pitfalls of differing risk attitudes and risk sharing in deriving an interval group utility function by transforming an ordinal scale, it is necessary to start with the assumption that each individual's preferences are measured on an interval scale. This approach has been developed primarily by Keeney (Keeney, 1975; Keeney and Kirkwood, 1975; and Keeney and Raiffa, in press). Keeney (1975) assumed that all individual utility functions, u_i, and the group utility function, u, satisfy the expected utility axioms, and showed that there are procedures for aggregating the individual utility functions into a group utility function that satisfy the Arrow conditions stated in terms of interval utilities rather than ordinal preferences. The critical property that must be met is:

$$\frac{\partial u}{\partial u_i} > 0, \quad i = 1, 2, \ldots, N.$$

One aggregation procedure that satisfies this condition is the weighted sum of the individual utility functions, that is,

$$u = \sum_{i=1}^{N} k_i u_i, \quad k_i > 0, \quad i = 1, 2, \ldots, N.$$

Keeney demonstrated how this group utility function satisfies the five conditions. A sufficient condition for the existence of a group utility func-

tion of this form is: In situations where the utilities of N-2 of the individuals are fixed for two alternatives, the group should be indifferent if, individually, each of the other two individuals is indifferent (Keeney, 1975). This assumption is similar to one from multiattribute utility theory, termed by various authors marginality, marginal equivalence, value independence, or additive independence that was first shown by Fishburn (1965) to be a necessary and sufficient condition for the existence of an additive multiattribute utility function.

Keeney (1975) has taken other results from multiattribute utility theory and reinterpreted them in the group utility context to arrive at conditions for the existence of other forms for aggregating individual utility functions into a group utility function. For example, the multiplicative form will be appropriate if an assumption is met that is slightly weaker than the one for the weighted sum procedure. The assumption is: In situations where the utilities of N-2 of the individuals are fixed for two alternatives, the utilities of the two remaining individuals shall guide the group decisions. The group utility function is then expressed by

$$u = \sum_{i=1}^{N} k_i u_i + k \sum_{\substack{i=1 \\ j>i}}^{N} k_i k_j u_i u_j + \ldots + k^{N-1} k_1 k_2 \ldots k_N u_1 u_2 \ldots u_N,$$

or the equivalent form,

$$ku + 1 = \prod_{i=1}^{N} (k k_i u_i + 1),$$

where $k_i > 0$ for all i and k is the solution to

$$k + 1 = \prod_{i=1}^{N} (k\, k_i + 1).$$

These results may at first be surprising since the theorem proved by Arrow for ordinal preferences must necessarily hold for preferences measured on a stronger scale, for example, cardinal. This is true because any cardinal scale is also an ordinal scale, although the converse is not true. Therefore, cardinal utility scales, include all of the properties of ordinal preference scales including those that allowed Arrow to prove his theorem. Arrow's formulation excluded the possibility of using individual strength of preference and interpersonal comparisons of utility. The Keeney formulations include both. Since strength of preference as measured by cardinal utilities alone is not enough to alleviate the paradox, the use of interpersonal comparisons of utility must be what allows the fulfillment of the five conditions. This raises a problem for which there is no very satisfactory solution. How can this comparison be made? There is good reason to believe that single individuals can make the necessary judgments about the relative differences in utility for themselves of

various alternatives, but who can say that the utility of one alternative for one individual is more or less than the utility of another alternative for another individual? Yet this comparison must be explicitly made. For example, in the weighted sum and multiplicative aggregation procedures discussed above, the k_i's take on this role. The larger k_i is, the more important the i^{th} individual's utility is in determining the group utility. Kirkwood (1972) discusses some of the problems and attempted solutions to assessing these weights.

Another factor to be considered in using these aggregation procedures is that they may result in decisions which are in some sense "unfair" or inequitable. For example, the weighted sum utility function for a two-person group with $k_1 = k_2$ implies that the group should be indifferent between the following alternatives (Keeney and Kirkwood, 1975):

A. $u_1 = 1$ and $u_2 = 0$.

B. A 50-50 chance of either $u_1 = 1$ and $u_2 = 0$, or $u_1 = 0$ and $u_2 = 1$.

C. A 50-50 chance for either $u_1 = 1$ and $u_2 = 1$, or $u_1 = 0$ and $u_2 = 0$.

Even assuming that equal utilities have the same meaning for both individuals, alternative A appears unfair since individual 2 has no chance of receiving his or her preferred outcome, while in each of alternatives B and C, each individual has an equal chance of receiving his or her preferred outcome. In addition, alternative C seems to be more acceptable than B, because no matter what happens, each individual will receive equivalent outcomes. Such considerations might lead to rejection of the weighted sum as an aggregation procedure.

Now consider the same alternatives when the group utility function is multiplicative with $k_1 = k_2 = 4$ and, therefore, $k = 1.25$. The group utilities for the alternatives are .4, .4 and .5, respectively. Although alternatives A and B are still indifferent for the group, alternative C is preferred to both. This seems to be a more equitable aggregation procedure.

Keeney and Raiffa (in press) point out another problem that can arise if the multiplicative aggregation rule is used. Add the following alternative to those already under consideration:

D. $u_1 = .48$ and $u_2 = .48$.

Under the multiplicative aggregation rule, this alternative has a utility of .43 for the group. Thus, it should be preferred to alternative B by the group. But notice that this violates Pareto optimality since the expected utility of alternative B is .50 for both individuals. Presumably, both individuals would prefer alternative B to alternative D while the group would prefer D to B.

II. D. Procedures for Reducing Disagreement

No formal procedure for aggregating individual preferences or utilities into a group preference or utility function seems to be completely satisfactory. But such aggregation processes must and do occur, if not formally, at least in some intuitive form. The question remains: Can we improve on these intuitive procedures? Some experimental work, but mostly experience, suggests some procedures which do not have an underlying formal structure may aid decision makers in this type of situation.

Experience suggests that simply providing more structure to the process of determining preferences often will reduce disagreement among members of groups. Gardiner (1974) provided some experimental evidence that this reduction in disagreement actually occurs. He found less disagreement when a highly structured multiattribute utility procedure was used to determine preferences than with simple holistic judgments. Procedures such as this not only may reduce disagreement, but will also help focus on the exact points of disagreement, which can then be considered specifically.

Another widely used procedure for reducing disagreement in assessing group preferences or values is called the Delphi procedure. Although it was developed and is used primarily for answering more factual questions, it can also be used to assess value judgments. There are many variations of the Delphi procedure, but all satisfy three general requirements: anonymity of the group members' responses, iteration with controlled feedback between rounds, and a statistical group response representing the group opinion or value. Since there can be no truth against which to compare value judgments, experimental work investigating the validity of the Delphi method in assessing value consists primarily of examining the reliability and the reduction in disagreement produced by this procedure.

The monograph by Dalkey, Rourke, Lewis and Snyder (1972) summarizes several experiments conducted at The Rand Corporation. The value judgments obtained using the Delphi method were generally found to be quite reliable and some convergence (reduction in disagreement) was found across rounds, mostly between the first and second rounds. However, the convergence was not as striking as the convergence typically obtained using Delphi to answer factual questions. Generalizations from Delphi findings using factual questions to value judgments are rather suspect so little is really known about how the Delphi procedure compares with other methods in actually reducing disagreement among group members making value judgments. The Delphi procedure in general is discussed more thoroughly in a later section of this paper dealing with group assessment of probabilities.

II. E. Summary

In concluding this discussion of procedures for assessing group utilities or preferences, the obvious conclusion to draw is that no entirely satisfactory method for deriving group utilities exists. All of the formal procedures for aggregating individual preferences or utilities into group preferences or utilities have some undersirable traits: restrictive applicability,

unfairness, violation of Pareto optimality, and so forth. The informal procedures suffer from the lack of experimental support and underlying theory. Experimental support for any of the methods is very difficult to obtain because of the validity problem. Since no "true" utility exists, a criterion for validity is difficult to define. Therefore, the crucial problem in assessing group utilities is not what procedure to use, but rather to develop satisfactory methods for validating assessed utilities.

III. INDIVIDUAL VERSUS GROUP JUDGMENTS OF UNCERTAIN VARIABLES

In this section, the distinction between groups that actually serve in a decision making capacity and groups that provide expert advice as input to another decision maker becomes important. Groups of the former type are not concerned with the individual-versus-group question, although it should have been considered prior to determining that a group rather than a single individual would function as the decision maker. However, in seeking expert advice, the decision maker should consider whether a group judgment is likely to be better than a single individual's judgment. This problem will be discussed in this section of the review, though many other considerations should also enter into the choice of individual versus group judgments, for example, costs and group size.

Several factors enter into the individual-versus-group question. How is the group judgment determined? What measure of goodness is used to compare individuals and groups? To what single individual should the group be compared? These questions must all be answered in order to make a judgment about whether group judgments will be better than individual judgments; and these qeustions will be considered subsequently in this review.

The relative merits of various procedures for determining the group judgment will be discussed in a subsequent section of this paper specifically for probabilistic judgments. Many studies that have compared individual with group judgments have used the arithmetic mean of the individuals in the group as the group judgment. This seems to be a natural choice since in our egalitarian society, it is probably representative of the way many groups actually function. In addition, the arithmetic mean has several desirable statistical properties, for example, less variance than individual judgments. Other methods sometimes used for determining a group judgment include median judgment, weighted arithmetic mean, geometric mean, and consensus (agreement on a single judgment through group discussion).

The measure of goodness used in comparing groups with individual depends entirely on the situation and type of judgment being made. When the true value that is being judged is known, individual and group judgments can be compared directly with the true value. However, for many types of judgments there is no known true value, for example, utility judgments, so other measures must be used. The most commonly used measure of goodness for probability judgments are proper scoring rules. (For theoretical developments, see Aczel and Pfanzagl, 1966; Toda, 1963. For more practical applications, see Murphy and Winkler, 1970; Stael von Holstein, 1970; Winkler, 1967).

In determining whether to use a group or a single individual judgment, the group judgment would be used if it were known a priori that the group judgment would be better than the best individual judgment. Similarly, an individual judgment would be used if all individual judgments were known to be better than the group judgment. Such knowledge is quite unlikely to be available, however, so the group judgments are typically compared with the judgment of the average individual. This comparison has some intuitive appeal since there is usually little or no rationale for a priori

judging the relative competency of the group members.

This section of the review begins with a discussion of some of the social psychological literature on group judgments. Much of this literature is quite old, and some is not directly relevant to group assessment of probabilities and utilities. Yet is does form part of the underlying rationale for using groups rather than individuals. A good further source on the topic of group performance is Davis (1969). He reviews group performance on decision making tasks as well as problem solving and learning from a perspective similar to that taken in this review. The primary concern is the product of the group and not the group structure or process by which the product is formed.

III. A. Statisticized Groups

Lorge, Fox, Davitz and Brenner (1958), in their review of individual versus group performance, draw several conclusions from the early experimental work comparing individual judgments with those of "statisticized" groups that seem quite relevant to consideration of group judgments of probability and utility. Statisticized groups are not actual face-to-face or interacting groups, but rather are groups formed so that a statistical procedure (usually averaging) can be used to obtain a group judgment from individuals making their own judgments. Beginning with the first recognized use of statisticized groups by Knight in 1921, reported by Lorge, et al., different results are apparent when two broad categories of stimuli are used. For numeric judgments of factual stimuli such as room temperature (Farnsworth and Williams, 1936; Knight, 1921), weight (Bruce, 1935; Gordon, 1924; Stroop, 1932), numerosity of buckshot (Bruce, 1935), and numerosity of beans (Klugman, 1945), statisticized groups outperform the average individual. However, the results are not so clearcut when the stimuli are more ambiguous and value-laden. Knight (1921) found no difference between groups and individual judging intelligence from the pictures of children; and Smith (1931) found some improvement in group as compared with individual judgments of personality and behavior traits from written reports, but not as much as was found in other studies of numeric judgment. Eysenck (1939), using another type of value judgment, found substantial increases in correlations between groups and the "expert" judgment over individual correlations. This result is not surprising, however, since the "expert" judgment was defined as the average judgment of 700 students and the groups were composed of samples from those same students' judgments.

The judgments required for factual and value-laden stimuli seem somewhat similar to the judgments required for probabilities and utilities. Probability judgments are in some sense factual since they typically are confirmable at some point in the future. However, utility judgments are value judgments and, therefore, seem to be more similar to judgments about intelligence, personality, and so forth. Such a generalization is, of course, quite weak; but to the extent that it is true, this research implies that group judgments of probability are more viable than group judgments of utility.

This type of research, comparing statisticized groups with individuals, has been subjected to considerable criticism. One such criticism is that the results are statistical artifacts. Stroop (1932) repeated Gordon's (1924) experiment with judgments of weight and included a second condition where a single individual made multiple judgments. The results were the same for groups composed both of judgments from different individuals and of many judgments from the same individual, leading Stroop to be the first to argue that such results simply demonstrate the statistical principle of error reduction and not any psychological principle of group processes. Zajonc (1962), in fact, showed that the superiority of statisticized groups over individuals can be predicted quite accurately by analytic techniques.

Although this criticism is quite damning from a social psychological point of view, from the decision analytic perspective adopted in this paper, these results remain useful. Certainly, reduced error is a desirable trait in the judgments necessary for decision making regardless of how it is achieved. What must be known are the characteristics of stimuli and situations that lead to this error reduction.

A standard result from test theory suggests that the heterogeneity of the group may affect the quality of the statisticized group judgment. Reinterpreting the equation for the validity of a test as a function of the test length (Gulliksen, 1950; Nunnally, 1967) leads to the following equation for determining the validity of the group judgment:

$$r_{t\bar{x}} = \frac{\sqrt{k}\ \bar{r}_{tx_i}}{\sqrt{1 + (k - 1)\ \bar{r}_{x_i x_j}}},$$

where $r_{t\bar{x}}$ is the correlation between the true value and the mean individual judgment; $\bar{r}_{tx_i}$ is the average correlation between the true value and individual judgments (averaged over all individuals); $\bar{r}_{x_i x_j}$ is the average intercorrelation among individuals; and k is the number of individuals. Considering $\bar{r}_{x_i x_j}$ as a measure of homogeneity, the validity of the statisticized group judgment clearly increases as homogeneity decreases, all other things being equal.

The empirical results of Jenness (1932) support this argument. Subjects made individual estimates of the number of beans in a bottle (true value = 811) and were subsequently assigned to three-person groups. The groups then discussed the judgments and reached a consensus for a group judgment. Subsequently, each individual again made an independent judgment. Jenness compared four types of groups: groups formed to maximize the diversity of original judgments of group members; groups formed to

minimize the diversity of opinion; groups formed "naturally" for discussion within a classroom; and a control group with no discussion. The results showed that for the diverse and natural groups, while the accuracy of the average individual did not increase, the majority of the subjects did improve their accuracy following the discussion. This result did not hold for the minimal diversity groups, nor was there any improvement simply from reconsidering the judgments a second time (control group). In addition, the average group error was less than the average individual error both before and after discussion. Thus, although diversity of opinion does not appear to be a necessary condition for group judgment to be superior to individual judgments, it does seem to lead to improved judgments from many individuals when such diversity is recognized through discussion.

There is also a large body of research in social psychology that suggests that the superiority of group performance over individual performance may not be due simply to the reduction in error variance. A theory has been developed known as social facilitation which explains the results of many studies that showed individual performance improved in the presence of other individuals (Zajonc, 1965). However, some tasks have also led to deteriorated performance. Zajonc indicated that there is a single feature that distinguishes between tasks on which individual performance improves in the presence of others and those on which the performance deteriorates: Improvement occurs on tasks that are well-learned, while deterioration takes place when new behavior is being learned. This research suggests that if we are to take full advantage of group judgment, the tasks being performed by the group must be well-learned.

The familiarity of the stimuli may also affect the differences in individual and statisticized group judgments. Farnsworth and Williams (1936) and Klugman (1945) investigated statisticized group judgments for what each termed "unfamiliar" stimuli, although their choices of "unfamiliar" stimuli are quite different. Klugman compared judgments of the numerosity of jacks and marbles (familiar) and lima beans and marrow beans (unfamiliar) and found that the statisticized group judgment was better than the average individual for the unfamiliar stimuli, but was no different for the familiar stimuli.

The unfamiliar stimulus for weight judgments used by Farnsworth and Williams was a box that had been constructed to take advantage of the size-weight illusion. Subjects hefted two boxes and then estimated the weight of the specially constructed box. The group judgments were no closer to the true weight than individual judgments, a result seemingly contrary to Klugman's results. However, given the types of stimuli used, this is certainly not surprising. What these experiments showed probably had little to do with whether the stimuli were familiar or not, but rather they illustrated that the statisticized group judgments were not better than individual judgments when all the individuals had a similar bias. The process of averaging reduces the error variance, not a constant bias.

Recently Einhorn, Hogarth, and Klempner (1975) have provided some

analytic support for this argument. They began by assuming that there is a population of individuals from which judgments can be obtained and that the distribution of the individual judgments is normally distributed. The true value of the quantity being judged does not necessarily coincide with the mean of the population of individual judgments, that is, some bias may exist. They then compared mean judgments of various size groups with the judgment of a randomly selected individual using expected utility for the comparison. Utility was assumed to be a linear function of the absolute difference of the judgment and the true value. Expected utilities for these two strategies were calculated for biases (difference between the population mean and the true value) varying from zero to three standard deviations, and group sizes from two to sixteen. On the average, the mean of a group always outperformed a single individual regardless of the bias. However, as the bias increased, the difference between the two strategies decreased rapidly to virtually no difference. This would seem to explain the lack of difference between groups and individuals in the Farnsworth and Williams (1936) experiment where an extensive bias presumably was present.

Einhorn et al. also examined another strategy for obtaining judgments which they called the "best-person strategy." If the person whose judgment will be best can be selected with certainty, their results indicated his or her judgment should be used. Using expected utility for comparing the judgment of the best person with the group mean judgment, the best-person strategy was shown to be superior to the group mean for all group sizes and all degrees of bias.

Recognizing the impossibility of identifying a single best person in most practical situations, Einhorn et al. also examined a more interesting case where the best person could be identified only with some probability less than one. Specifically, they assumed the rank order of individual judgments was known and that any individual's probability of being identified as the best person was inversely proportional to the rank. For example, in a three-person group, the best person (rank = 1)would have a probability of 3/(3+2+1)= .50 of being selected as the best person. Similarly, the second best person would have a probability of .33 of being chosen as best, and the third best person would be chosen as the best person with probability .17. The group judgments were then calculated as a weighted sum of the individual's judgments; the weights being the above defined probabilities.

For the low range of bias, that is, biases up to .7 standard deviation of the distribution of individual judgments, the group mean did better. However, when the bias exceeded .7, the weighted sum with weights inversely proportional to rank did better than the group mean, suggesting that in situations where there is a possibility of extensive bias, effort to identify best persons may be more fruitful than simply using a group mean.

Einhorn et al. also argued that the proportional weighting scheme is conservative since it does not yield weights differing greatly from equal weights and, therefore, the probability of correctly identifying the best person is actually larger than suggested by this scheme. This implies that the best person strategy may be relatively better than is suggested by their results.

The empirical evidence bearing on the question of whether the superior members of a group can be identified a priori is equivocal at best. Such attempts have been numerous and varied, but no general pattern of results has emerged. Jenness (1932) found no correlation between intelligence and errors in judging the number of beans in a bottle. Kaplan, Skogstad and Girshick (1950) found correlations of .60 between scores on a test of social problems and success in predicting social events, and between scores on a science test and success in predicting scientific events. The actual differences in success of prediction between low and high scores on the tests were, however, rather small.

Many recent attempts to identify group members whose judgments can be expected to be superior have focused on self-ratings as a measure of expertise. Experimental efforts to improve the Delphi procedure through the use of individual self-ratings have had limited success. Studies at The Rand Corporation using the Delphi procedure to obtain group judgments about general information questions, the answers to which might typically be found in almanacs, showed error in the group judgments tended to decrease as the average self-rating of the group members increased (Brown and Helmer, 1964; Dalkey, 1969a; Dalkey, Brown and Cochran, 1970b). They did not, however, find any relationship between individual self-ratings and individual error. More recently, Brockhoff (1975) found no association between either individual or group self-ratings and error in factual and forecasting judgments on economic questions made by bankers.

This evidence suggests that even if some members of a group have the ability to make better judgments than result from the group procedure, such individuals will be difficult or impossible to identify.

With this background on the general use of statisticized group judgments, we now turn to the specific topic of using statisticized groups for assessing probabilities. The scope of this research is quite limited, and much work is needed to determine if the findings for statisticized group judgments in general carry over to this specific type of judgment.

III. B. Statisticized Group Judgments in Probabilistic Forecasting

Several attempts to improve probabilistic forecasting procedures have focused on the aggregation of individually assessed probabilities into a group probability usually using averaging as the aggregation procedure. Aggregated group probabilities are then compared to the individually assessed probabilities using a proper scoring rule. Since the understanding of these scoring rules is crucial in explaining some results, a slight digression into the meaning and use of proper scoring rules seems to be in order (for a more general discussion see Stael von Holstein, 1970; or Winkler, 1967).

The defining characteristic of all proper scoring rules is that if an assessor has $\underline{p}$ as his true subjective probability distribution, and reports $\underline{r}$ as his probability distribution, his expected score on a proper scoring rule will be maximized if and only if $\underline{r} = \underline{p}$. Thus, proper scoring rules are used in eliciting probability judgments to reward the assessor for honesty,

that is, for reporting his true beliefs.

There are three commonly used proper scoring rules:

the quadratic scoring rule,

$$Q_k = 2r_k - \sum_{j=1}^{n} r_j^2 \quad ;$$

the spherical scoring rule,

$$S_k = r_k \Big/ \sqrt{\sum_{j=1}^{n} r_j^2} \quad ;$$

and the logarithmic scoring rule,

$$L_k = \log r_k \quad ;$$

where Q_k, S_k, and L_k are the scores obtained if the k^{th} event occurs, r_j is the subjective probability assigned to the j^{th} event, and r_k is the probability assigned to the event that occurs.

Proper scoring rules can also be used to evaluate forecasts. If $\underline{p}$ is the "true" probability distribution, then the expected score on a scoring rule is maximized if $\underline{r} = \underline{p}$; that is, the probability distribution given by the assessor is equal to the "true" distribution.

This technique has been used to compare the scores of statisticized group forecasts with the average score of individual forecasts on tasks including meteorological prediction (Stael von Holstein, 1971a), scores of football games (Winkler, 1971), stock prices (Stael von Holstein, 1972), general information questions (Gough, 1975), and short term socio-economic predictions (Brown, 1973). The results of these comparisons were all quite similar: the statisticized groups always outperformed the average individual. This should come as no surprise since the quadratic and logarithmic scoring rules used in these studies are concave functions on the probability simplex as are all strictly proper scoring rules. Because of this property, the score of an average of individual probabilities must be better than the average of the individuals' scores. In fact, Brown (1973) has shown that the difference between the quadratic score of the average probability distribution and the average of the individual scores can be determined simply by knowing the individual probability distributions without knowledge of the actual outcome of the event being predicted.

There is, however, further evidence, which does not depend on proper scoring rules for evaluation, that suggests the statisticized group probability judgments are superior to individual judgments. For example, Winkler (1971), in addition to evaluating probability assessments with scoring rules, also observed the outcomes of hypothetical bets on the football games.

The bets were made by comparing the expected point spreads expressed

by the subjects with the spread given by bookmakers, and assuming the subjects would take the side of the bet that was favorable from their point of view. The expected point spreads assessed by the subjects were reasonable approximations of the means of the assessed probability distributions. Three betting schemes were considered: bet one dollar on each game, bet the absolute value of the difference between the subject's spread and the bookmaker's spread, or bet the squared difference of the subject's and bookmaker's spread. For each betting scheme the mean judgment of the group outperformed the average individual on both Big Ten Conference and National Football League games in terms of money won or lost. The difference ranged from 2¢ to 47¢ for every dollar bet, with the 47¢ difference on the squared difference betting scheme for Big Ten games. This type of economic evaluation is a valuable contribution to the comparison of individual and group probabilistic judgments since it does not suffer from the limitations of the proper scoring rules.

Dalkey (1975) has developed additional theoretical arguments that suggest group probability judgments are generally more satisfactory than individual judgments based on economic evaluations. Whereas Winkler evaluated the assessments in terms of objective expectation, for example, with knowledge of actual outcomes, Dalkey examined subjective expectations in the following situation: Consider a group faced with a decision as specified by a decision matrix. The group must choose a single course of action, with the action chosen depending on the assessed probabilities of possible events. The payoff to each individual is proportional to the proceeds realized by the group. Under these circumstances, the average subjective expectation of the individual group members will be less than or equal to the expected payoff based on the average probability distribution.

Dalkey points out that such a procedure may not satisfy group members since each will believe that the group would have done at least as well or better if it had chosen a course of action based on his or her own probabilities. This leads to consideration of another payoff strategy in which each group member is paid proportionally to what the group would have made if it had followed the individual's own advice. In this situation, the average expectation of the total group can be maximized by each member's adopting the average group estimate as his or her own. Thus, "almost any way you view an enterprise, if there is disagreement on probabilities or utilities, but agreement on the rule of common action, the expectation of the group judgment is greater than the average expectation of the individuals" (Dalkey, 1975, p. 22).

Although Dalkey's arguments are quite persuasive, they must be interpreted with some caution. The expectations considered are all subjective expectations, and the well-documented existence of biases in subjective probabilities suggests that the objective expectations may be quite different from the subjective expectations. The work by Einhorn et al. (1975) in particular suggested that we must consider the existence of biases in probability assessment in comparing group and individual judgments.

One of the biases that has been found in assessing subjective probabilities is what has been referred to as "overconfidence." That is, the subjective estimates express more certainty than is warranted by the objective relative frequencies. This bias is evident in both discrete and continuous subjective probability distributions (Lichtenstein, Fischhoff, and Phillips, 1976).

Amos Tversky and Daniel Kahneman have studied other types of biases. (See Tversky and Kahneman, 1974, for a review of their work). They propose three heuristics that people use in making judgments about uncertainty: representativeness, availability, and anchoring and adjustment. These heuristics are useful in simplifying the information processing necessary to make probabilistic judgments, but can often lead to systematic errors.

Representativeness is a heuristic often used in judging the probability that a specific instance belongs to a general class. One form of this heuristic is the disregard subjects have for sample size when judging the probability of a sample statistic. For example, the probability of more than sixty percent of the births at a hospital being male on any given day was judged by most subjects to be the same regardless of the average number of daily births (Kahneman and Tversky, 1972). Subjects seem to consider only the degree to which the sample proportion represents the population.

The availability heuristic suggests that people judge the probability of the occurrence of an event by the ease with which they can remember other instances of that event. For example, people judge the letter k more likely to be the first rather than the third letter in a word, although the converse is actually true (Tversky and Kahneman, 1973). Words in which the first letter is k are easier to recall than words in which the third letter is k.

The final heuristic proposed by Tversky and Kahneman is called anchoring and adjustment. The general idea of this heuristic is that when asked to make a numerical judgment, people first anchor on some seemingly appropriate starting point and then adjust usually insufficiently away from the anchor. This heuristic has been demonstrated in subjects' judgments of unknown percentages. A percentage was chosen by a random device and the subjects indicated whether the true percentage was above or below the selected percentage. Then the subjects made a numerical estimate of the true percentage. The lower the randomly selected percentage (the anchor), the lower the subjects' estimated percentage.

The preceding discussion of biases in probability assessment is by no means a complete or comprehensive treatment of the subject. Rather it is meant to illustrate the widespread existence of such biases in individual judgments, and the possibility of similar biases in group judgments.

When a relatively large bias exists in the judgments of all group members, Einhorn et al. (1975) suggest that attempting to identify the better or best members of the group will lead to better group judgments than will a simple averaging process. Will this type of strategy be useful in obtaining group probability judgments? Existing empirical evidence implies probably

not. The data comparing the performance of individual subjects with an "average subject" using proper scoring rules as the performance criterion indicate that very few subjects outperformed the average. In fact in the studies where these data are available (Brown, 1973; Stael von Holstein, 1971a, 1972; Winkler, 1971), only 4% to 13% of the individual subjects outperformed the group judgment formed by taking the average of the individual judgments.

Stael von Holstein (1972) provided a vivid example of the superiority of a group probability judgment over individual judgments with the group judgment being a simple average of the individual members' judgments. Using five groups of subjects, bankers, stock market experts, statisticians, business administration teachers, and students of business administration, he found that the average judgment of the group of stock market experts outperformed every individual (n=98) based on average scores from the quadratic scoring rule. Similar results were obtained in an earlier study (Stael von Holstein, 1971a) using meteorologists, research assistants in meteorology, statisticians,and meteorology students to predict rainfall and temperature. In this study the prediction formed by taking the average of the individual judgments of the research assistants outperformed all but one individual (n=30) again based on the quadratic score. These results certainly must be taken as an argument against attempting to identify the "best" individual judge. They do point to another strategy that might be fruitful, namely to attempt to identify a "best" group. If relatively reliable procedures can be found that will predict the quality of the judgments of groups, their use could be quite rewarding.

III. C. Summary

Reviewing the work comparing individual and group judgments leads to the following conclusions:

1. In general the group judgment will be more accurate than individual judgments primarily due to a decrease in error variance.

2. The superiority of group judgments over individual judgments may be greater for factual judgments than for value judgments.

3. A larger diversity of individual opinion among group members will lead to greater superiority of the group judgment over individual judgments.

4. The task being performed by the group should be well-learned.

5. The existence of similar judgmental biases in most or all group members will reduce the superiority of the group judgment.

6. When such biases exist, the "best-person strategy" (Einhorn et al. (1975) may be superior to a statisticized group judgment.

7. The a priori identification of "best" individual judges will be difficult or impossible.

8. A statisticized group probability assessment will always be superior to the average performance of individual group members as evaluated by a proper scoring rule due to the mathematical properties of the scoring rule.

9. If a group is faced with a decision making problem, the expected utility associated with the use of a group probability judgment (average of individual judgments) is greater than the average expectation of the individuals.

Thus, in most situations where some type of judgment is required for decision making, group judgments should be preferred to individual judgments if obtaining group judgments costs more.

IV. PROCEDURES FOR FORMING PROBABILISTIC GROUP JUDGMENTS

The preceding section concluded that in general group judgments will be superior to individual judgments. But how should such group judgments be obtained? Most of the comparisons of individual and group judgments in the preceding section contrasted individual judgments with group judgments formed by averaging the individual judgments. An extension of this procedure, weighted linear combinations, is discussed first, along with several other mathematical procedures for aggregating individual probability assessments into a group assessment. The next section focuses on behavioral approaches to forming the group judgments. These behavioral approaches to varying degrees depend on some type of communication or interaction between individual group members hopefully to increase the agreement among the individuals. In some cases complete agreement or consensus may result in a single judgment that can be used to represent the total group opinion. In other cases the behavioral procedure will not necessarily produce a consensus so one of the mathematical aggregation procedures must be used in conjunction with the behavioral method.

IV. A. Mathematical Aggregation Procedures

The procedures to be discussed in this section have a common approach to determining a group probability estimate from individual estimates. Each uses some mathematical procedure to combine the individual distributions into a single distribution that is meant to represent in some sense a consensus of the opinion of all the individuals. The individuals involved are not necessarily a group in the usual meaning of the word because there is no requirement that they actually meet. The individuals form a group only in that there is more than one individual.

These procedures may vary in the mathematics of the aggregation, the formal justification, and the exact type of judgment required from the individuals as input to the procedure. Evaluation of the usefulness of each procedure depends on ease of application, justification, and performance in the sense that the product of one procedure may be generally better as measured by some criterion, say proper scoring rules.

IV. A. 1. Weighted linear combinations. Since simple averages of individual judgments generally are superior to individual judgments, a logical approach to improving these simple averages would be to use a weighted average, which gives higher weights to "better" judges. This brings back the question of whether such judges can be identified a priori, and if so, is the differentially weighted combination better than the equally weighted one.

This approach, termed the "opinion pool" by Stone (1961), has been tried experimentally by several investigators using many different weighting procedures usually based on one of two measures of quality: self-ratings or performance. Stael von Holstein (1972) conducted the most extensive investigation of differential weighting procedures. Two different consensus groups were used; a fixed group consisting of the nine Ss with the best average quadratic scores in the first half of the study, and a variable group, changing from session to session, composed of the nine Ss who did best in

the preceding session. Eight weighting procedures were used: (1) equal weights; (2) weights proportional to average scores for preceding sessions; (3) weights proportional to scores in the preceding session; (4) weights proportional to moving averages with the latest score having a relatively small weight; (5) weights proportional to moving averages with the latest score having a relatively high weight; (6) weights proportional to ten minus the rank of the individuals' scores for previous sessions; (7) weights proportional to the inverse of the rank; and (8) weights proportional to the participant number. Comparing the quadratic scores of the two consensus groups and the weighting procedures showed relatively little difference in the weighting procedures, with procedures 6 and 7 doing slightly better in the fixed group and 3 doing slightly better in the variable group. The variable group in general outperformed the fixed-composition group.

Other studies comparing different weighting systems based on self-ratings and prior performance evaluated by proper scoring rules have led to similar results: virtually no difference between the group judgments resulting from the different weighting systems (Gough, 1975; Stael von Holstein, 1971a; Winkler, 1971). In addition to self-ratings, Rowse, Gustafson, and Ludke (1974) used peer ratings in their study in which firemen judged the relative likelihood of the occurrence of various fire and ambulance alarms. Using the difference in the logarithm of the estimated odds and the logarithm of the true odds as the evaluation criterion, the use of weights based on peer ratings in various combinations with self-ratings again led to basically the same results.

A point should be made here about the use of weights based directly on scores of previous sessions. Such weights are measured on a ratio scale while the scores are measured on only an interval scale. That is, if R is a proper scoring rule, then any positive linear transformation of R, $aR + b$, $a > 0$, is also a proper scoring rule. Obviously, making such a transformation with $b \neq 0$ will affect the resulting weights. As Stael von Holstein (1971a 1972) points out, the choice of the exact form of the proper scoring rule will greatly affect the relative weights given to the best and worst group members. Even such arbitrary choices seem to have little effect on the quality of the group judgment (Stael von Holstein, 1971a).

The above weighting procedures are based on heuristic schemes. Roberts (1965) and DeGroot (1974) offer more formal procedures for determining weights. Roberts presented a Bayesian procedure for updating weights as data are collected. Prior weights are assigned to group members and are then revised by multiplying the prior weights by the predictive probability placed on the data by the individual judge. The predictive probability $f(x)$ is the expected value of the likelihood in light of the prior distribution:

$$f(x) = \int_{\Theta} h(x,\theta)d\theta = \int_{\Theta} g(\theta)f(x|\theta)d\theta \quad ;$$

where $h(x, \theta)$ is the joint distribution of the data and the parameters θ, $g(\theta)$ is the prior distribution of θ, and $f(x|\theta)$ is the likelihood of the data given θ. That is, the weight assigned to individual i after some data have been observed, w_i', is proportional to the product of the prior weight assigned to i, w_i, and $f_i(x)$:

$$w_i' \sim w_i f_i(x).$$

Winkler (1971) encountered some difficulty applying this weighting procedure to produce group judgment. Ten subjects were predicting the point spreads of football games on a weekly basis. Using uniform prior weights, the posterior weights were then proportional to the product of the probabilities assigned to the outcomes that actually occurred. Thus, any subject assessing a probability of zero for the outcome that occurred was assigned a weight of zero in calculating all future group judgments. Even assigning several near-zero probabilities to such outcomes eventually led to weights of zero, with weights given only to four decimal places. At the end of thirteen weeks of predicting, only two of the original ten subjects had non-zero weights. The judgments formed using this weighting scheme had lower average scores on the quadratic and spherical scoring rules, but higher scores on the logarithmic scoring rule. Winkler attributed the difference in results for the different scoring rules to the fact that the logarithmic scoring rule is the only proper scoring rule consistent with the use of Roberts' weighting procedure. That is, the evaluation of probability assessors using likelihood ratios as suggested by Roberts and proper scoring rules will be the same if and only if the scoring rule used is the logarithmic (Winkler, 1969).

To avoid the problem of a high percentage of zero weights, Winkler tried setting any probabilities less than .01 equal to .01. This resulted in a third subject with a non-zero weight and a slight drop in scores on all scoring rules. A cutoff of .10 was also tried, leaving six of the subjects with non-zero weights, but lowering the scores even more.

DeGroot (1974) formulated the weighting problem in the following manner: Consider how any individual group member might revise his or her subjective probability distribution upon learning the distributions of the other group members. One possibility is for the revised distribution to be a linear combination of the distributions of the other group members. Let p_{ij} be the weight assigned by individual i to the distribution of individual j. Also assume that $p_{ij} \geq 0$ for all i and j and $\sum_{j=1}^{k} p_{ij} = 1$, for all i, where k is the number of group members. This revision may be considered an iterative process. Each individual finding that all other members may also have revised their distributions again revises his or her own distribution using the same set of weights as before. Now with $\underline{P}$ being the k-by-k matrix of weights p_{ij} and $\underline{F}$ the k dimensional vector of individual distributions with transpose $\underline{F}' = (F_1, \ldots, F_k)$,

$$\underline{F}^{(n)} = \underline{P}\,\underline{F}^{(n-1)} = \underline{P}^n\,\underline{F}$$

where $\underline{F}^{(n)}$ is the vector of individual distributions after n iterations. The question is under what conditions all k components of $\underline{F}^{(n)}$ will converge to the same limit as $n \to \infty$. Defining p_{ij}^n as the elements of $\underline{P}$, this convergence will occur if and only if there exists a vector $\underline{p}^* = (p_1^*,\ldots,p_n^*)$ such that

$$\lim_{n\to\infty} p_{ij}^n = p_j^*$$

for all i and j. Using the theory of Markov chains, DeGroot proves that $\underline{p}^*$ exists if there exists a positive integer, n, such that every element in at least one column of $\underline{P}^n$ is positive (Theorem 1, p. 119). The most obvious situation satisfying this requirement is one in which at least one member of the group receives non-zero weights from all other members. The vector $\underline{p}^*$ can be calculated by solving the set of linear equations $\underline{p}^*\underline{P} = \underline{p}^*$ subject to the constraint

$$\sum_{i=1}^{k} p_i^* = 1.$$

The crucial assumption in the development of DeGroot's weighting procedure is that the group members revise their distributions as linear combinations of the distributions of the other members of the group. While this assumption may have some intuitive appeal, it does not have any normative rationale, nor is there any evidence that it is descriptive of revision in such a situation. These same arguments, however, must apply to all weighted averaging procedures for forming group judgments.

The other assumption made by DeGroot that might be questioned is that the weights assigned to other group members do not change on subsequent iterations. Group members may very well want to change the weights they assign as the distributions of other members are revealed. For example, if person A found the distribution of person B, to whom he had assigned a non-zero weight, was not consistent with facts that A knew to be true, A certainly might want to reduce or eliminate the weight assigned to B.

These weighting procedures suggested by Roberts and DeGroot certainly deserve further empirical testing. DeGroot's procedure has not been tested at all, while Roberts' procedure has had only one rather unsuccessful application. However, given the similarity of results using different heuristic weighting methods, the use of these procedures would be expected to produce results that differ little from other weighting procedures.

In summary, the weighting procedure used to form a statisticized group probability judgment makes little or no difference in the quality of the resulting judgment. This should really come as no surprise given knowledge of the insensitivity of linear models to changes in their parameters (von Winterfeldt & Edwards, 1973). Relatively large changes in the input parameters of a model will have only a small effect on the output of the model. This phenomenon has been demonstrated elsewhere for regression models (Dawes and Corrigan, 1974; Einhorn and Hogarth, 1975).

The evaluation of weighting schemes suffers doubly from this insensitivity. First, the weighted linear combination of individual probability assessments will change relatively little even with relatively large changes in weights; and second, the scoring rule used to evaluate the resulting group judgment will be insensitive to changes in the group probability assessment. Given these theoretical arguments and the experimental evidence, simplicity argues for use of equal weights. Why bother trying to determine a weighting procedure if there is little chance of improving results?

IV. A. 2. The pari-mutuel method. Several methods other than weighted linear combinations have been suggested as procedures for aggregating individual probability judgments into a group probability judgment. These procedures are generally more complex, yet have a better theoretical basis than the linear combination method.

One of the first mathematical solutions to the problem of forming a consensus from individual probability assessments was proposed by Eisenberg and Gale (1959). Recognizing the existence of an institution that performed exactly this function, the pari-mutuel betting system used at race tracks, they suggested a similar approach for generally obtaining consensus probability distributions. The pari-mutuel model is formulated as follows: Assume there are m individuals and n possible mutually exclusive events. Let p_{ij} be the probability that individual i places on event j. Each individual has an amount, b_i, to bet and bets so as to maximize his subjective expectation. The final consensus probabilities, $q_1,\ldots,q_n$, are proportional to the amount bet on each event. Individual i will maximize his expectation by betting only on those events for which the ratio p_{ij}/q_j is maximum.

The model appears to be somewhat circular at this point. The individual must know the consensus probabilities in order to know on which events to bet, while the consensus probabilities cannot be determined until all individuals have made their bets. The question answered by Eisenberg and Gale is whether or not final consensus probabilities and individual bets exist that are compatible with the pari-mutuel system and the individual strategy of maximizing expectation. They proved that these probabilities and bets do exist and in fact the consensus probabilities are unique.

To prove this a function F with mn arguments is defined:

$$F(x_{11},\ldots,x_{mn}) = \sum_{i=1}^{m} b_i \log \sum_{j=1}^{n} p_{ij}x_{ij},$$

with $x_{ij} \geq 0$ for all i, j, and $\sum_{i=1}^{m} x_{ij} = 1$. Solving this equation for the values $(\bar{x}_{11},\ldots,\bar{x}_{mn})$ that maximize F allows the computation of the consensus probabilities. These probabilities are given by:

$$q_i = \max_i \frac{\partial F}{\partial \bar{x}_{ij}} = \max_i \frac{b_i p_{ij}}{\sum_{j=1}^{n} p_{ij} \bar{x}_{ij}} .$$

The amount individual i bets on alternative j, β_{ij} is

$$\beta_{ij} = \bar{x}_{ij} q_j .$$

From this equation we see that $\sum_{j=1}^{n} p_{ij} \bar{x}_{ij} = \sum_{j=1}^{n} \beta_{ij} (p_{ij}/q_j)$,

which is the expectation of individual i. Thus, the function F shows that the group is maximizing the sum of logarithms of individual expectations weighted by the amount the individual has to bet.

Although this model is formally rather appealing, in some instances the consensus probabilities resulting from its application are quite non-intuitive. For example, Eisenberg and Gale point out that for two individuals with equal amounts to bet and two possible events, if either individual assigns a subjective probability of .50 to both events, the consensus will assign probabilities of .50 to both alternatives regardless of the probabilities of the second individual.

Norvig (1967) extended the pari-mutuel model and provided a more satisfying mathematical formulation of the procedure by which the final consensus probabilities are determined. He formulated the procedure as a dynamic iterative process in which individuals place bets that determine a set of consensus probabilities, which, in turn lead the individuals to place new bets, and so on. He then proved that the consensus probabilities converge to the same probabilities specified in the Eisenberg-Gale model formulation as the number of iterations increases.

The pari-mutuel model is relatively well-known and often referenced in the literature on group probability assessment, yet it has never been tried either experimentally or in actual applications. It certainly deserves some empirical testing, especially to compare the probabilities determined by the pari-mutuel model with those arrived at by other models. It is also of particular importance, given the possibility of non-intuitive results mentioned above, to ascertain whether pari-mutuel probabilities are acceptable to a decision making group.

IV. A. 3. Aggregation using conjugate distribution. Another procedure for combining individual probability judgments into a group judgment based on conjugate distributions and the use of Bayes' Theorem has been suggested by Winkler (1968). A distribution is said to be a member of a conjugate family of distributions if the posterior distribution, arrived at by applying Bayes' Theorem to revise the prior distribution in light of an observed sample, and

the prior distribution are members of the same family of distributions. For example, if the prior distribution is a beta distribution and the sample observed is generated by a Bernoulli process, the posterior distribution will also be a beta distribution. Conjugate distributions are useful since they reduce the application of Bayes' Theorem to very simple arithmetic. The Bernoulli data and beta prior distribution with parameters α and β will yield a posterior beta distribution with parameters α' equal to the sum of α and the number of observed "successes" in the data and β' equal to the sum of β and the number of observed "failures." DeGroot (1970, Ch.9) provides a useful discussion of conjugate distributions along with examples of most conjugate families.

The use of conjugate distributions proposed by Winkler assumes the subjective distributions of all group members are members of the same family of conjugate distributions. The group probability distribution is determined by successive applications of Bayes' Theorem using all the individually assessed distributions. Each individual's probability distributions may be weighted according to some desirable characteristic (see previous section) to specify the individual's contribution to the group distribution.

The assignment of these weights poses a problem in addition to determining the relative weight of each individual: there is no constraint that the weight must sum to one as there was for the linear combination procedures, so the sum must also be specified. Winkler advanced an argument that the range of the sum of the weights should be restricted to between one and k, where k is the number of individuals in the group. This seems to be a reasonable restriction if two extreme cases are considered. In the first case, all group members are completely independent, that is, their judgments are based on completely different and independent information. Since the information provided by each individual does not overlap with information provided by any other individual, all weights would be one and the sum would be k. At the other extreme is the case where all individuals use the same information to assess their probability distributions. Each individual should then assess the same distribution, and the group distribution would be identical to the individual distribution. In this case the sum of weights would be one.

Winkler compared the distributions formed using the conjugate procedure and the weighted linear combination procedure. He assumed there were two group members and the probability distribution of each was a beta distribution. The comparisons were made with various relative weights for the two individuals and different combinations of parameters for the individual distributions. The weights for the conjugate procedure always summed to one. The most notable difference in the two procedures was that the conjugate procedure usually produced a much tighter distribution. This occurred because in addition to the assigned weights, the process by which the group distribution is formed using conjugate distributions naturally weights each individual according to the tightness of the individual distribution. The tighter the individual distribution, the more similar the group distribution is to the individual's distribution. This characteristic, considered in con-

junction with the bias toward assessing distributions that are tighter than justified by the individual's knowledge (Alpert and Raiffa, 1969; Seaver, von Winterfeldt and Edwards, 1975; Stael von Holstein, 1971b), may lead to a situation where more weight is being given to individuals with greater biases. The group distribution will become even tighter if the weights assigned in the conjugate procedure are allowed to sum to more than one. The distributions considered by Winkler are in fact the least tight distributions that can be produced if the sum of the weights is constrained to between one and k.

Each of these two combination procedures has other favorable and unfavorable characteristics. Even if all individual distributions are unimodal, the group distribution may be multimodal if a linear combination is used. The conjugate procedure will always produce a unimodal distribution in this case. The multimodal distribution may simply be unacceptable in many decision making situations, making the use of a linear combination of individual distributions undesirable. On the other hand, the conjugate procedure requires some judgment as to the independence of the distributions assessed by the various individuals in order to determine what the sum of the weights should be. In addition, use of the conjugate procedures constrains the distributions of all individuals to membership in the same family of distributions, a restriction that often may not be justified. The wide variety of conjugate families and permissible parameters within families does allow considerable diversity in the allowable distributions. For example, the normal, gamma, and beta distributions will probably provide adequate approximations for most univariate distributions with no, one, or two bounds, respectively, on the range of the variable being considered. If the conjugate method is to be used effectively, procedures need to be developed for determining the family of distributions that is most applicable in a given case and assessing the parameters of the individual distributions.

Winkler and Cummings (1972) investigated the use of the linear combination and conjugate methods in an experiment where subjects were provided individual probability distributions and asked to determine a consensus distribution to use in decision making. The stimuli were stock prices one year in the future broken down into six intervals. The individual distributions were from a normal family, and a self-rating of expertise was presented with each individual distribution. For half the stimuli, the subjects simply chose one of four consensus distributions, calculated using the linear combination and conjugate procedures with equal weights and weights proportional to self-ratings. All weights summed to one. For the remaining stimuli, subjects generated their own distributions given the individual distributions and self-ratings.

The results showed that subjects generally preferred to use distributions formed by linear combinations. On the questions offering a choice of distributions, 71% of the chosen distributions were linear combinations. Linear combinations also provided the best fit (least squares) to 76% of the subjectively generated distributions. For both types of judgments and both combination procedures, the distributions chosen or generated by the subjects utilized weights proportional to self-ratings in a large majority of the cases.

Thus, the linear combination procedure appeared to be a more accurate indication, descriptively, of the way in which people integrate several probability distributions into a single distribution. This is not really surprising since averaging is a familiar and simple mathematical process. This experiment said nothing, however, about the normative validity of the two methods. Which procedure provides a more valid representation of the uncertainty associated with an unknown quantity? Does one procedure generally lead to decisions with a higher expected utility? These normative questions remain to be answered. The evidence presented above indicates that the answers to these questions may depend primarily on the acceptability of the multimodal distributions produced by the linear combination method and the tightness of the distributions produced by the conjugate procedure.

IV. A. 4. The expert-use model. Another approach to aggregating several individual probability distributions into a single distribution has been developed by Morris (1971, 1974, 1975) within the larger general context of modeling the use of experts in decision analysis. His elegant treatment of the problem is philosophically and mathematically consistent with the Bayesian approach to decision making. For exactly these reasons, this model is a highly desirable step in the right direction toward solution of the problem of resolving disagreement among experts. Although this approach seems usable in the cases of a single expert or multiple independent experts, it becomes intractable for dependent multiple experts, the situation with which this review is primarily concerned.

In Morris's model, the individually assessed probability distributions are multiplied by a calibration function (Morris, 1975) and then treated as likelihood functions and combined via Bayes' Theorem to produce a posterior or aggregate distribution. The calibration function is used to eliminate any known biases in the expert's probability assessment and may be determined either subjectively or on the basis of previously collected data. In the multiple-expert case, the function that represents the individual assessments of all the experts, called a surrogate prior by Morris, is simply the normalized product of the individually assessed distributions and a joint calibration function. The joint calibration reflects both the individual calibration of each individual and the interdependency of the individual assessments. This is the function that is extremely difficult to ascertain in practice unless the individual distributions are independent, in which case it is simply the product of the individual calibration functions.

Some similarities between this model of expert use and the conjugate procedure (Winkler, 1968) are apparent. The expert-use model is more general in that it does not require that all individual distributions be members of the same conjugate family. The expert-use model also does not allow the explicit assignment of weights to individual distributions as does the conjugate procedure. Since the expert-use model explicitly deals with the dependence of individual assessments, in general the results from the use of these two procedures will not be the same. In fact even if the individual distributions are all members of the same family of distributions, the surrogate prior will be a member of the same family only if the joint calibration function is also a member of that family, a rather unlikely circumstance.

IV. A. 5. The probabilistic approach. Dalkey (1975) has devised an approach to the aggregation problem that he calls the probabilistic approach which is similar to Morris's model of expert-use in that it is based on the use of Bayes' Theorem and an explicit expression of the dependence of the set of individual probability assessments. The primary difference is that Dalkey works with estimates of $P(E_i|R_j)$, the probability of event i occurring given the subjective distribution of R_j of individual j, while Morris uses estimates of $P(R_j|E_i)$. The $P(E_i|R_j)$ estimates can be obtained from realism curves; functions that relate assessed probabilities to the relative frequency of occurring events. For instance, of all the events to which an assessor with perfect realism assigns a probability of .70, approximately 70% should actually occur. Estimates of $P(R_j|E_i)$ do come into play in Dalkey's formulation of the dependency term in his equation for calculating the group probabilities from the individually assessed probabilities. The dependence of the joint set of individual assessments, R, with respect to a particular event, E_i, is defined as

$$D_R^{E_i} = \frac{P(R|E_i)}{\prod_{j=1}^{m} P(R_j|E_i)},$$

where $P(R|E_i) = P(R_1, R_2, \ldots, R_n|E_i)$. That is, the dependence reflects the difference between the probability of the joint occurrence of a set of individual assessments and the product of the probabilities of the individual assessments.

The basic equation for determining the aggregated probability of each of the mutually exhaustive events E_k is

$$P(E_k|R) = \frac{\prod_{j=1}^{n} P(E_k|R_j)}{\sum_{i=1}^{m} D_{ik} \prod_{j=1}^{n} P(E_i\ R_j)},$$

where

$$D_{ik} = \frac{D_R^{E_i}}{D_R^{E_k}} \left(\frac{U(E_k)}{U(E_i)} \right)^{n-1}$$

The $U(E_k)$'s are the prior probabilities of the events based on whatever information is available prior to knowing the R_j's. In many cases the priors may be uniform, that is, all equal to 1/m.

Like Morris's model, this model is usable only if the individual distributions are independent. An impasse is reached when the distributions are dependent because of the difficulty in determining D_{ik}.

Although this procedure may not be applicable in practice, Dalkey used the results to examine the conditions under which the group assessment would do better than the individual assessments. Using the logarithmic scoring rule to measure performance and defining the net score as the difference between the score for an assessment (group or individual) and the score for a uniform distribution, Dalkey showed that the group score is equal to about n times the average individual score plus a function of the dependency terms. Thus, if the dependency terms are small, and the average individual net score is positive, the group does about n times better than the average individual. On the other hand, with the small dependency terms, the group will do n times worse if the average individual net score is negative. The most favorable condition for the group assessment was also shown not to be complete independence, but negative dependence, in which case the joint probability of the group assessment $P(R_1,R_2,\ldots,R_n)$ is less than the product of the probabilities of the individually assessed distributions $P(R_1)P(R_2)\ldots P(R_n)$.

IV. A. 6. Summary. The preceding discussion of mathematical aggregation procedures leads to the conclusion that the procedures which are better justified theoretically, for example, the expert-use model and the probabilistic approach, are difficult or impossible to apply in practice. The least theoretically justified procedure, the weighted linear combination method, is the easiest and most widely applied approach to the combination problem. The only competition to the linear combination method comes from the pari-mutuel model and conjugate procedure. The pari-mutuel model is applicable only for discrete variables. However, continuous variables can, of course, be approximated very well by breaking them down into discrete categories. The conjugate method suffers from the problems of determining the expected sum of the weights and requiring that all individually assessed distributions be members of the same family. These two approaches need further empirical testing to compare their appropriateness and ease of application with the linear combination method.

The usefulness of the expert-use model and the probabilistic approach comes from their explication of the problem of specifying the dependence of individual assessments and the role this dependence should play in determining the group probability distribution. Obviously, investigation of methods for specifying the dependence can be a fruitful topic for future research.

IV. B. Behavioral Approaches

As noted in the preceding discussion, mathematical aggregation as a means of producing a group decision does not require that the individuals actually form a group in the usual meaning of the word. However, there is a wealth of research that shows the superiority of actual groups to individuals in problem solving, decision making, and other similar types of tasks (Collins and Guetzkow, 1964; Davis, 1966; Maier, 1967). Explanations have included many factors contributing to this superiority: increased available information, increased reliability, increased acceptance of group product, social facilitation, distribution of responsibility, and so forth. While some of these characteristics can be achieved by statisticized groups, other require communication and interaction among group members.

A study by Holloman and Hendick (1972) suggested that the more interaction among group members, the better the quality of the resulting decisions. They studied the decision making of six-person groups with six different procedures for reaching a group decision: (1) average of individual decisions; (2) decision by a leader chosen by the group; (3) decision by a two-person committee chosen by the group; (4) majority vote after discussion; (5) consensus through discussion; (6) consensus after a majority vote. The decision faced by the groups resulted from viewing the film Twelve Angry Men, which depicts the deliberations of a jury in a murder trial, particularly the personalities and interactions among the jury members. The first vote by the jury was 11 to 1 for conviction. Prior to the second vote, the film was stopped and the groups were presented with the decision problem: one-by-one, the jury members changed their votes to not guilty. What is the order in which they change? The results showed decreasing error (group decision compared with the true outcome of the film) from procedure (1) to procedure (6). Since procedures (1) to (6) appear to lie on a continuum from less interaction to more interaction, this result suggested that increasing the interaction of group members improved decision making.

This study can be criticized on several points; for example, the criterion for determining the quality of the decisions was not necessarily a normative criterion since it was only the judgment of the person who wrote the story for the film. It does, however, suggest that the amount of interaction among group members should be investigated as a possible means of improving the judgments made by groups for decision making.

The primary contribution of the possible superiority of interacting groups over non-interacting groups probably comes from the increased information available to the group members. Obviously, the group as a whole has at least as much information available to it as any single individual has, and certainly in most cases substantially more. If each group member can increase the information available on which to make the required judgments by interacting and communicating with other group members, presumably the quality of the judgments should improve. This transmission of information between group members can take place only through some means of communication which is not utilized by statisticized groups.

An additional factor that should be considered in comparing interacing and non-interacting groups is the reaction of the group as a whole to the group product. This becomes particularly important when the group is involved in real decision making. The group judgments must be acceptable so that the group will be agreeable to basing important decisions on the judgments. Even if the group is willing to accept maximization of expected utility as a decision rule, it cannot use decision analysis if the group judgments of probabilities and utilities are not trusted. In discussing behavioral approaches to determining group judgments, this factor should be kept in mind.

Although several factors suggest that interacting groups may be superior to non-interacting groups, the social psychological research also shows that such interaction may produce results that are detrimental to performance. For example, interacting groups may spend considerable time and effort on structuring the group. Other considerations that may interfere with the group's

attaining maximum efficiency include the presence of dominant personalities; status incongruity among group members; and pressure for conformity, that is, reaching a consensus may become more important than the quality of the actual product. Collins and Guetzkow (1964) summarize the research on these and other obstacles to efficient group decision making, while Van de Ven and Delbecq (1971) consider similar factors inhibiting group problem solving. In order to effectively use interacting groups, methods need to be found that minimize the inhibiting influences without eliminating the beneficial aspects.

The approach taken in dealing with this problem has been to restrict the interaction and communication among group members. Hopefully, such an approach will allow the interaction that is necessary for the facilitating factors to function while eliminating or at least reducing the interaction leading to poorer performance.

Although the types of restrictions that might be tried are innumerable, research seems to have focused on two basic procedures for restricting the interaction: the Delphi method, developed by Norman Dalkey, Olaf Helmer and their associates at The Rand Corporation; and the use of "nominal" groups developed by Andre Delbecq, Andrew Van de Ven and their associates at the University of Wisconsin.

As these techniques are described in the following subsections of this review, you will note that use of the Delphi and nominal group procedures does not produce a consensus of group opinion. Here, and in what follows, consensus is defined as general agreement among group members on the final product of the group itself. The key here is that the group itself determines the product. Groups using both the Delphi and nominal group techniques do not actually determine the product themselves. Rather, some mathematical technique is still necessary to combine the individual judgments into a group product. Thus, these techniques may be viewed as a combination of the usual consensus group procedures and the mathematical aggregation methods discussed in the previous section of this paper.

The Delphi method and the nominal group procedure are described, and some of the relevant research is discussed in the following subsections of this review. Since these procedures were not developed specifically for assessing subjective probabilities, there has been relatively little research on the use of these techniques to derive group probabilities. The studies that have been done on this topic are reviewed in the final subsection.

IV. B. 1. The Delphi method. One of the difficulties encountered in reviewing the literature on the Delphi technique is the broad range of procedures that have adopted this name. Because of the widespread use of this procedure in dealing with real-world prediction problems, some practitioners apparently feel that simply using this name lends credence to their investigations. Perhaps there is also a certain sex appeal in the name that contributes to the extensive use and abuse of the name and methodology.

Nevertheless, there are three necessary features that were defined by the

originators of Delphi and seem to be characteristic of many so-called Delphi investigations. They are: (1) anonymity of group members; (2) iteration with controlled feedback; (3) statistical group response (Dalkey, 1969b). The anonymity and lack of direct interaction among Delphi group members serve to eliminate the effects of dominant personalities, status incongruities, pressure for conformity, and so forth, while the use of controlled feedback on successive rounds allows the exchange of ideas and information.

A typical Delphi exercise will have several individuals making the requested judgments anonymously, usually by questionnaire, although recent technological developments make data collection via computer networks feasible (Linstone and Turoff, 1975, Ch. VII). The Delphi manager then summarizes the first round results in some manner and feeds them back to the individuals who make another set of (possibly different) judgments based on the new information. This process can be continued for any number of rounds (usually two or four). After the final round, some statistic, for example, average or median, is used as the single judgment representing the group opinion.

Delphi has been used extensively in both the private and public sectors as a tool for forecasting. This literature is not discussed here. Extensive annotated bibliographies are included in Pill (1971) and Sackman (1974) and the recent publication edited by Linstone and Turoff (1975) contains a complete discussion of the Delphi method and related techniques and a comprehensive bibliography. This paper will concentrate on the experimental work comparing Delphi with other procedures designed for the same purpose and on the development of specific procedures to improve the general Delphi method.

Dalkey and Helmer (1963) reported the first use of the Delphi method in a 1951 experiment attempting to elicit expert judgments about the number of A-bombs that would be needed to reduce the U.S. munitions output to a certain level. They concluded that the procedure was successful in reducing the disagreement among the participating experts, but many problems remained. One major problem was the validity of the resulting answer. Reduction in disagreement is certainly a legitimate goal, but even more important is the correctness of the final group judgment. In this case, there was no known correct answer against which to compare the judgment of the group of experts.

The validity question was attacked by Dalkey and his colleagues in a series of experiments in the late 1960's (Dalkey, 1969a, 1969b; Dalkey, Brown and Cochran, 1970a, 1970b). In these experiments, college students were used as subjects, and the questions they were answering were general information questions selected from almanacs. Such questions have characteristics similar to the prediction questions that Delphi was designed to answer: the true answer is unknown to the Delphi participant, but the participant has some information relevant to the question. In addition, almanac questions have the useful property that the true answers are known to the experimenter so validity can be assessed.

The first question addressed in these studies was the comparison of the Delphi method with face-to-face discussion. The results of two small experiments slightly supported the superiority of the Delphi method (Dalkey, 1969b). In the first experiment, twenty almanac questions were given to two groups,

each with five college students. One group was asked to reach a consensus via face-to-face discussion, while the second group used the Delphi method with four rounds of estimates, feedback consisting of the medians and quartiles of the estimates on the preceding round, and the medians of the fourth round estimates taken as the group judgment. The Delphi group was more accurate on 13 questions; the discussion group was superior on 7 questions.

Additional support for the Delphi method came from the second study where medians and quartiles were fed back between rounds one and two, and face-to-face discussion was used between rounds two and three. There was slightly more improvement between rounds one and two than between two and three.

This was the extent of the comparison of Delphi and face-to-face groups reported in the Rand experimental work. These results are an extremely weak base for assuming the superiority of the Delphi method, yet further experimental work focused on procedures for improving the Delphi method rather than on building a better underpinning for the superiority of the method.

The series of studies reported by Dalkey (1969a, 1969b) also investigated the amount of convergence and improvement over rounds in the Delphi groups. Both convergence and improvement were obtained, although the convergence was much more striking than the improvement. Between rounds one and two, the group response (median of individual responses) improved on 64% of the questions where a change occurred. The results also showed that the judgments seemed to be over-converging; "the increase in accuracy is not commensurate with the reduction in spread" (Dalkey, 1969b).

The improvement across rounds shown in the group responses in this series of experiments was encouraging but hardly spectacular. Therefore, subsequent studies investigated mechanisms for increasing the improvement. Two approaches to this question were studied extensively; variations in the type and amount of feedback between rounds, and use of self-ratings to identify subgroups that would tend to be more accurate. Dalkey (1969a) reported that feedback of medians and quartiles produced improvement, while simply re-estimating on multiple rounds did not. Feedback between the second and third rounds of reasons for judgments outside the interquartile range in addition to medians and quartiles did not produce any change in improvement. Another type of feedback, each individual's percentile with respect to the entire group, also resulted in improvement similar to that from feedback of medians and quartiles (Dalkey, et al., 1970a). Although feedback does improve the judgments, the type of feedback apparently has little effect.

The results of attempts to select more accurate subgroups on the basis of self-ratings are more conflicting. Brown and Helmer (1964) were successful in identifying such subgroups while Dalkey (1969a) was not. This problem was further investigated by Dalkey et al. (1970b) who succeeded in forming more accurate subgroups on the basis of self-ratings provided two conditions were met: the size of the subgroup must be substantial (operationalized as at least seven), and the difference in self-ratings between high and low self-ratings groups must be substantial (operationalized as a difference of

at least one point on the five-point rating scale between the highest member of the low self-rating group and the lowest member of the high group.

The results of this series of studies at Rand have been interpreted as supporting the superiority of Delphi procedures over traditional methods for obtaining group judgments. However, the similarity of the judgments required for answering almanac questions and for predicting actual unknown events has been justifiably questioned. Are the same cognitive processes actually used for both types of judgments and, therefore, are the results generalizable? The validity problem is difficult to handle when the "true value" is actually unknown, a characteristic of real-world prediction situations.

To show generalizability of the Delphi method to judgments, where the true value is actually unknown at the time the judgment is made, Dalkey and Brown (1971) had subjects answer short-term prediction questions where the true answer would not be known for one to three months. The group answers to these prediction questions were found to improve at least as much with feedback and re-estimation as answers to almanac questions. These results certainly support the generalizability of the Delphi method. They do not, however, confirm the validity of the method for long-range prediction, since there are still possible basic differences between short- and long-range forecasting; and the subjects used in this study were college students, not real-world experts.

A recent experiment by Brockhoff (1975) has also compared performance on almanac and forecasting questions by groups of experts. This experiment used actual experts (bankers) to answer both almanac questions and forecasting questions in their field of expertise (economics) using both Delphi methods and traditional face-to-face discussion groups. The Delphi groups tended to have less error on the almanac questions than did face-to-face discussion groups, but the reverse was true for forecasting questions where the face-to-face discussion groups outperformed the Delphi groups. Additionally, the Delphi groups performed better as a whole on the almanac questions than on the forecasting questions, while the face-to-face discussion groups performed equally well on both types of questions. Furthermore, the lack of significant correlations between individual performance on the two types of questions indicated a basic difference in the skills needed for answering these questions. These experimental results certainly do not support the use of the Delphi method as a forecasting technique and in fact conflict with the results obtained in the Rand experiments.

In summarizing the experimental work comparing Delphi methods with other procedures for determining group judgments, no definite conclusions can be drawn. Little experimental evidence is available making this comparison, and much of what is available is based on college students answering almanac questions. The more realistic experimental work is conflicting, with the most realistic study (Brockhoff, 1975) resulting in generally negative findings with respect to the Delphi method. Potential users of the Delphi method should be aware of the lack of experimental support and approach applications with caution. Obviously, more experimental work is needed. However, the inaccessibility of real experts and the problem of validating any judgments other than

answers to almanac or short-term prediction questions make such research difficult.

Other reviewers have additional criticisms of the Delphi method. Weaver (1972) has argued that the primary value of the method is as a learning device rather than as a forecasting method. The judgments required for the Delphi exercise may stimulate the participants to more critical thinking about the problem and lead to increased understanding of the complexities involved. Such an argument is probably equally true of other procedures designed to serve the same purpose as Delphi.

Pill (1971) has suggested that the value of the Delphi method is as a communication device rather than as a subjective scaling technique. The evidence presented above indicates that even this may be an optimistic appraisal. Pill also emphasized that Delphi should be viewed as a utilitarian approach to quantifying opinion and as such should be judged on the results it produces rather than on any theoretical underpinning. This paper adopts a similar rationale in evaluating the Delphi method as a procedure for quantifying group judgment.

In contrast, Sackman (1974), the severest of published Delphi critics, attacked the methodology on the basis of a lack of underlying theory. He cosidered the Delphi procedure as a form of psychological test and illustrated how the methodology failed to meet the standards for psychological tests. Further criticisms by Sackman included the process by which "experts" are selected, lack of a true consensus, ambiguity in questions, and the lack of empirically demonstrated validity. He concluded that "conventional Delphi is basically an unreliable and scientifically unvalidated technique in principle and probably in practice" (p. iv). Although many of Sackman's criticisms are well-founded, to completely drop the use of the Delphi method at this point seems to be throwing out the proverbial baby with the bath water.

IV. B. 2. The nominal group method. Generally, researchers referring to groups as nominal groups mean a collection of people in physical proximity with no spontaneous interaction among the group members. For example, research involving nominal groups will typically have several people together in a room working on the same problem, but with no communication between the people. However, in this paper, the nominal group method refers to a much more specific technique, developed by Delbecq and Van de Ven (1971) for purposes similar to those leading to the development of the Delphi method: to take advantages of the known superiority of group processes while eliminating the detrimental effects.

Van de Ven and Delbecq (1971) reviewed much of the literature on the effectiveness of nominal versus interacting groups for problem solving, and concluded a process based primarily on nominal groups, but incorporating a certain amount of interactive discussion was the best approach to use in problem solving groups. The particular approach adopted, well-described in Delbecq, Van de Ven, and Gustafson (1974), is a four-step procedure: (1) silent judgments by individuals in the presence of the group; (2) presentation

to the group of all individual judgments without discussion; (3) group discussion of each judgment for clarification and evaluation; (4) individual reconsideration of judgments and mathematical combination.

This process differs from the Delphi method in two distinct features: judgments are made in the presence of other group members so, therefore, members are not anonymous; and direct communication is allowed between group members, rather than written feedback provided by the Delphi manager. An additional difference that may not be obvious from descriptions of the two methods is the role of the leader for nominal groups. No leader is necessary for groups using the Delphi method, but the nominal group leader may play a very important role in nominal groups, particularly at step (3) where he or she must guide the discussion and evaluation. At this step, it is critical that the leader control discussion to keep it focused on the issues so as to minimize personality effects. Probably the most effective type of leader in this situation is a "distant" leader (Fiedler, 1960; 1967). This type of leader is predominantly goal-oriented with little interest in interpersonal relations with other group members. Research has shown that a task with a well-defined structure, such as nominal group tasks typically have, is conducive to success with this type of leadership (Shaw, 1971). Delbecq et al. (1975) provide an excellent discussion of the advantages and disadvantages including the leadership role of both the nominal group method and the Delphi method and thorough descriptions of the actual use of both procedures.

An experimental comparison of groups using the Delphi method, the nominal group method, and interacting groups was reported by Van de Ven and Delbecq (1974). The problem facing the groups was to develop a job description for dormitory counselors at a midwestern university. Sixty seven-person groups composed of mixtures of students, housing administrators, faculty, and academic administrators were created: twenty groups used each of the three methods. Two criteria of effectiveness were used: quantity of ideas and perceived satisfaction of group members.

In terms of quantity of ideas, the nominal groups were slightly (not significantly) more effective than the Delphi groups, while both were significantly more effective than the interacting groups. The measure of effectiveness, perceived satisfaction, which perhaps is more important in this experiment, was significantly higher for nominal groups than for Delphi and interacting groups which had virtually the same level of satisfaction. The importance of this measure is that it should reflect the willingness of the group to accept the results. Presumably the more satisfied the group is with the process by which a judgment (or set of judgments) is reached, the more satisfied it is with the resulting judgments. On the basis of these results, Van de Ven and Delbecq concluded that the nominal group method is generally the superior method for group decision making. The Delphi method was also shown to be superior to interacting groups, and in certain situations, it can profitably be used instead of the nominal group method.

In deciding between using the nominal-group method and the Delphi method, characteristics in addition to their relative effectiveness must be considered. For example, if the group members are geographically distant, there may be no

choice but to use the Delphi method. However, the Delphi method is a much slower process than the nominal-group method, since it involves several rounds of questionnaires. Turnaround time may be quite lengthy using the mail. Recent developments in the use of computers may eliminate this criticism of the Delphi method.

IV. B. 3. Experimental comparisons with probabilistic judgments. Both nominal group method and the Delphi method are relatively recent developments in group decision making procedures. In addition, neither was developed expressly for use by groups making probabilistic judgments. Therefore, it is not surprising that there has been little research on the effectiveness of these procedures in producing probabilistic judgments about uncertain events. In fact, to date only three experiments have been performed on this topic.

The first study (Gustafson, Shukla, Delbecq, and Walster, 1973) compared four procedures for determining a group judgment of the relative likelihood of two hypotheses. The hypotheses under consideration were whether a given height was more likely to be male or female. The four types of groups compared were statisticized groups, modified interacting groups, modified Delphi groups, and nominal groups. The modification in the interacting groups was that the group itself did not have to reach a consensus about the judgment. Rather, following discussion, each group member made his or her own individual estimate, and the individual judgments were subsequently mathematically aggregated to determine the group judgment. The only difference between the modified interacting groups and the nominal groups was that in the nominal groups, each group member made an individual judgment prior to group discussion. The Delphi procedure used was a modification of the normal Delphi method in that the group members were all present together as judgments were being made.

All judgments were made on a logarithmic scale of odds. In each of the four conditions, the group judgment was taken as a geometric mean of individual judgments of the four group members. Using the average deviation of the group likelihood ratios from the true likelihood ratios as the measure of goodness, the nominal groups produced the best estimates and the modified Delphi groups produced the worst estimates. The statisticized groups were only slightly better than the interacting groups.

A second study by Gough (1975) generally confirmed the Gustafson et. al. findings. Four group procedures were used with subjects making probabilistic judgments about almanac questions. The probabilistic judgments were encoded as five fractiles of the cumulative subjective probability distribution. Each procedure had subjects make individual judgments both before and after the group procedure. The procedures used by the groups were simple reconsideration with no information exchange, a Delphi method with written feedback between individual judgments, a procedure where the group had to reach a consensus between individual judgments, and a nominal group with verbal interaction but no necessary consensus. The group opinion was determined by averaging the post-treatment individual judgments, and the quadratic scoring rule was applied to evaluate the resulting group probabilities. Again in this

study, the nominal groups attained the best performance, followed by the consensus, Delphi, and reconsideration groups in that order. The same ranking was obtained in comparing the improvement between the pre- and post-treatment individual judgments supporting the hypothesis that the difference was due to the procedure used by the group. Gough did not report how the actual consensus probabilities determined by the consensus groups compared with the averaged post-treatment judgments. Since the groups using a verbal exchange of information outperformed the other groups in this experiment and the nominal groups were also the most effective in the Gustafson et al. study, these findings tend to disconfirm the hypothesis suggested by the developers of the Delphi method that verbal interaction should be avoided.

The final study was done by Fischer (1975) with the same four types of groups used by Gustafson et al., except that true consensus groups were substituted for the modified interacting groups. In this study, the judgments required were predictions of the freshman GPA of ten randomly selected members of a recent Duke University class. The subjects were given the gender, high school GPA, SAT math, and SAT verbal scores of the ten students as information on which to make the prediction. The predictions took the form of probabilities assigned to four ranges of GPA. A logarithmic scoring rule was used both to motivate the subjects (pay was based on the score) and to evaluate the group probabilities. The results of the application of the logarithmic scoring rule indicated there was virtually no difference between the four types of groups, thus failing to replicate the Gustafson et al. and Gough results. Fischer was apparently unaware of the Gough study, but suggested the difference in the Gustafson et al. results might be well due to the method of evaluation. Fischer evaluated the judgments stated in probabilities while in the Gustafson, et al. study, the judgments were stated and evaluated in odds. An example given by Fischer explains how large differences in odds may be insignificant in probabilities. In the Gustafson et al. study, one question had true odds of 1.8. The average judgment of the nominal groups for this stimuli was approximately 9.55, and the average judgment for the Delphi groups was approximately 19.05. While in the Gustafson et al. study, this is a rather large difference, if these odds are converted to probabilities, the nominal group judgment is .91 and the Delphi judgment is .95, an insignificant difference when compared with the difference of both from the true probability of .64.

Although this argument may account for the difference in the Fischer and Gustafson et al. studies, it does not account for the Gough results, which were evaluated in terms of probabilities. Two suggestions for future research can be gleaned from this research: more research is necessary and future studies should not depend on a single method for evaluating the probabilistic judgments produced by various group procedures.

Two additional studies deserve brief mention in the context of comparing behavioral consensus judgments with individuals or statisticized group judgments. Although neither study involved the use of Delphi or nominal groups, both explore the performance of actual consensus groups in making probabilistic judgments as a minor part of broader research interests.

In probabilistic weather forecasting, Stael von Holstein (1971a) found the consensus forecasts of two experienced weather forecasters had a slightly lower average quadratic score than the average forecast of the two forecasters. The consensus forecasts did outperform the average individual forecaster.

Goodman (1972) also compared the performance of consensus groups with individuals on likelihood ratio judgments of relatively abstract stimuli. Using the slope of the line regressing the logarithm of the estimated likelihood ratios on the logarithm of the true likelihood ratios as a measure of performance, she found that consensus groups did slightly better than the average individual. (In this case, "better" means a slope closer to 1.0, since that is the slope that would result from perfect performance).

These two studies add support to the hypothesis that groups will outperform individuals, even consensus groups which suffer from many detrimental effects.

IV. B. 4. Summary. The behavioral approaches to forming probabilistic group judgments in general seem to produce desirable results above and beyond those of statisticized groups. Many of the behavioral approaches, in fact, combine group interaction and communication with mathematical aggregation, hopefully to take advantage of the best of both methods.

Research comparing the various behavioral approaches certainly does not definitively single out any technique as superior. Delphi, the most widely used of the behavioral methods is also the most controversial. Experiments at Rand obtained generally favorable results, while other experimenters (Brockhoff, 1975) have been less enthusiastic. Other research has shown that actual face-to-face discussions improve results, contrary to a hypothesis underlying the development of the Delphi method. The nominal-group method had proved to be a particularly successful approach. In direct comparisons between the nominal-group method and the Delphi method, the nominal-group method has always done as well as Delphi and in most cases better.

Although the amount of research is obviously limited, and many questions remain to be fully explored, the nominal-group method seems to be the superior behavioral approach to forming group judgments based on current evidence.

V. CONCLUSION

Throughout this review, an implicit assumption has been made that groups should make decisions in accordance with the expected utility axioms as defined, for example, by von Neumann and Morgenstern (1947) or Savage (1954). These axioms are widely accepted as normative rules for decision making by individuals. Should they also be accepted for group decision making? This question of a normative theory of group decision making has not been addressed in this paper. The literature reviewed has been primarily oriented toward how the inputs necessary for making decisions can and should be determined when judgments can come from more than one individual. Yet ultimately what we really need is a theory of group decision making.

For most decision analysts, myself included, maximization of expected utility seems to be the foundation for a theory of group decision making. However, as illustrated in previous sections of this review, this may conflict with another compelling condition, Pareto optimality. How can such a conflict be resolved in a satisfactory theory of group decision making? One approach is simply to eliminate explicit consideration of the individuals within a group. The group becomes a superorganism that should satisfy the expected utility axioms. Individual probabilities and utilities are not considered nor, of course, is any relationship between individual and group parameters. The group-as-a-superorganism theory eliminates many of the problems discussed earlier in this paper. For example, the Arrow and Dalkey impossibility theorems become unimportant since there are no formal requirements on the relationship between individual and group preferences and probabilities. There can be no conflict between maximizing expected utility and Pareto optimality since individual expected utilities are never explicitly considered. At a theoretical level, considering the group as a superorganism is intuitively compelling. If a group actually functions as a decision maker, it certainly has its own identity. What is crucial for making decisions is the will of the group, not of single individuals. As is true for this normative theory applied to individual decision making, our concern is only that the preferences and opinions of the decision maker satisfy the axioms, not how these preferences and opinions are determined.

From a pratical viewpoint, the superorganism approach is much less compelling. In order to apply the theory, we must be concerned with how preferences and opinions are determined. If asked to, a group will probably be able to determine the utilities and probabilities necessary for calculating expected utilities. But if no instructions are given as to how these numbers are to be determined, the group will adopt its own procedure. As suggested in this review without guidance as to what type of a procedure to use, the group is likely to adopt a procedure that may produce suboptimal results. One of the purposes of this review is to provide information to decision analysts that they can use to guide the process by which a group determines the probabilities and utilities necessary for decision making.

In contrast to the superorganism theory, there is a practical approach to eliminating conflicts between maximization of expected utility and Pareo optimality that still utilizes individual probabilities and utilities. If the group

first considers probabilities and does not consider utilities until a probability distribution is agreed upon, many of the problems may be eliminated. Wilson (1968) has proved a series of theorems stating the conditions under which probability distributions and utility functions exist that represent the opinions and preferences of the group members and satisfy the necessary conditions for expected utility theory. One of the conditions is agreement on the probability distribution (Theorem 7, p. 129). That is, if all members of the group agree on the relevant probabilities, the group probability distributions and utility functions can exist for the group to use in decision making via maximization of expected utility.

Disagreement about probabilities seems to be fundamentally more resolvable than disagreement about utilities. Differences in probablity assessments depend primarily on differences in knowledge, available information, and perhaps biases. These differences can be recognized and often resolved, thereby leading to greater agreement about the disputed probabilities.

Differences in utilities are much less resolvable. Preferences are typically not based on knowledge or information, but rather follow from internal value systems. These personal values can be expected to differ among group members. The degree to which they differ will in part depend upon the purpose and goals of the group. For example, members of a corporate board of directors might all have relatively similar utilities since goals are generally well-defined. Presumably the primary goal is profit-making. There may be disagreement concerning other goals, but given their subordinate role to profit-making, these disagreements may be relatively minor. On the other hand, some groups may have such broad and general goals that group members may have extremely divergent utilities. Members of a city council may all agree that they want to do what is best for the city, but such a broad goal leaves considerable room for disagreement. Some council members may prefer a no-growth policy as the best means of achieving the goal while other members believe maximum growth is preferable. There is no reason to believe that such conflicting viewpoints can be resolved.

One final point about disagreement among group members should be made. For the group to effectively use expected utility theory as a decision making aid, group members do not necessarily have to agree on the probabilities and utilities. They need only agree on how group probabilities and utilities will be determined. If some reasonable rule for determining these parameters can be agreed upon prior to consideration of specific decisions, specific disagreements will be resolved by the agreed-upon rule.

VI. REFERENCES

Aczel, J., and Pfanzagl, J. Remarks on the measurement of subjective probability and information. Metrika, 1966, 2, 91-105.

Alpert, M., and Raiffa, H. A progress report on the training of probability assessors. Unpublished manuscript, Harvard University, 1969.

Arrow, K. Social Choice and Individual Values. New York: Wiley, 1951.

Arrow, K. Social Choice and Individual Values, 2nd Edition, New York: Wiley, 1963.

Bacharach, M. Group decisions in the face of differences of opinion. Management Science, 1975, 22, 182-191.

Black, D. The decisions of a committee using a special majority. Econometrica, 1948, 16, 245-261.

Brockhoff, K. The performance of forecasting groups in computer dialogue and face-to-face discussion. In Linstone, H., and Turoff, M. (Eds.), The Delphi Method: Techniques and Applications. Reading, Ma.: Addison-Wesley, 1975.

Brown, B., and Helmer, O. Improving the reliability of estimates obtained from a consensus of experts. Rand Paper P-2986, The Rand Corporation, Santa Monica, Ca., 1964.

Brown, T. An experiment in probabilistic forecasting. Rand Report R-944-ARPA, The Rand Corporation, Santa Monica, Ca., 1973.

Bruce, R. Group judgments in the field of lifted weights and visual discrimination. Journal of Psychology, 1935-1936, 1, 117-121.

Clark, R. Group induced shift toward risk: A critical appraisal. Psychological Bulletin, 1971, 76, 251-270.

Collins, B., and Guetzkow, H. A Social Psychology of Group Processes for Decision-making. New York: Wiley, 1964.

Coombs, C. Social choice and strength of preference. In Thrall, R., Coombs, C., and Davis, R. (Eds.), Decision Processes. New York: Wiley, 1954.

Dalkey, N. Analyses from a group opinion study. Futures, 1969, 1, 541-551. (a)

Dalkey, N. An experimental study of group opinion: The Delphi method. Futures, 1969, 1, 408-426. (b)

Dalkey, N. An impossibility theorem for group probability functions. Rand Paper P-4862, The Rand Corporation, Santa Monica, Ca., 1972.

Dalkey, N. Toward a theory of group estimation. In Linstone, H., and Turoff, M. (Eds.), The Delphi Method: Technology and Applications. Reading, Ma.: Addison-Wesley, 1975.

Dalkey, N. Group decision theory. Unpublished manuscript, UCLA, 1975.

Dalkey, N., and Brown, B. Comparison of group judgment techniques with short-range predictions and almanac questions. Rand Report R-678-ARPA, The Rand Corporation, Santa Monica, Ca., 1971.

Dalkey, N., Brown, B., and Cochran, S. The Delphi method IV: Effect of percentile feedback and feed-in of relevant facts. Rand Memorandum RM-6118-PR, The Rand Corporation, Santa Monica, Ca., 1970. (a)

Dalkey, N., Brown, B., and Cochran, S. Use of self-ratings to improve group estimates. Technological Forecasting, 1970, 1, 283-291. (b)

Dalkey, N., and Helmer, O. An experimental application of the Delphi method to the use of experts. Management Science, 1963, 9, 458-467.

Dalkey, N., Rourke, D., Lewis, R., and Snyder, D. Studies in the Quality of Life. Lexington, Ma.: Lexington Books, 1972.

Davis, J. Group Performance. Reading, Ma.: Addison-Wesley, 1969.

Dawes, R., and Corrigan, B. Linear models in decision making. Psychological Bulletin, 1974, 81, 95-106.

DeGroot, M. Optimal Statistical Decisions. New York: McGraw-Hill, 1970.

DeGroot, M. Reaching a consensus. Journal of the American Statistical Association, 1974, 69, 118-121.

Delbecq, A., and Van de Ven, A. A group process model for problem identification and program planning. Journal of Applied Behavioral Science, 1971, 7, 466-492.

Delbecq, A., Ven de Ven, A., and Gustafson, D. Group Techniques for Program Planning. Glenview, Il.: Scott, Foresman, 1975.

Einhorn, H., and Hogarth, R. Unit weighting schemes for decision making. Organizational Behavior and Human Performance, 1975, 13, 171-192.

Einhorn, H., Hogarth, R., and Klempner. E. When one head is better than two: A decision analysis. Unpublished manuscript, Graduate School of Business, University of Chicago, 1975.

Eisenberg, E., and Gale, D. Consensus of subjective probabilities: the pari-mutuel method. Annals of Mathematical Statistics, 1959, 30, 165-168.

Eysenck, H. The validity of judgments as a function of number of judges. Journal of Experimental Psychology, 1939, 25, 650-654.

Farnsworth, P., and Williams, W. The accuracy of the median and the mean of a group of judges. Journal of Social Psychology, 1936, 7, 237-239.

Fiedler, F. The leader's psychological distance and group effectiveness. In Cartwright, D., and Zandler, A. (Eds.), Group Dynamics: Research and Theory, 2nd Ed. Evanston, Il.: Row, Peterson, 1969.

Fiedler, F. A Theory of Leadership Effectiveness. New York: McGraw-Hill, 1967.

Fischer, G. An experimental study of four procedures for aggregating subjective probability assessments. Technical Report 75-7, Decisions and Designs, Inc., McLean, Va., 1975.

Fishburn, P. Independence in utility theory with whole product sets. Operations Research, 1965, 13, 28-45.

Fishburn, P. The irrationality of transitivity in social choice. Behavioral Science, 1970, 15, 119-123.

Fishburn, P. The Theory of Social Choice. Princeton, N.J.: Princeton University Press, 1973.

Gardiner, P.C. The application of decision technology and Monte Carlo simulation to multiple objective public policy decision making: A case study in California coastal zone management. Unpublished Ph.D. dissertation, University of Southern California, 1974.

Goodman, B. Action selection and likelihood ratio estimation by individuals and group. Organizational Behavior and Human Performance, 1972, 7, 121-141.

Gordon, K. Group judgments in the field of lifed weights. Journal of Experimental Psychology, 1924, 7, 389-400.

Gough, R. The effect of group format on aggregate subjective probability distributions. In Wendt, D., and Vlek, C. (Eds.), Utility Probability, and Human Decision-Making. Dordrecht-Holland: Reidel, 1975.

Gulliksen, H. Theory of Mental Tests. New York: Wiley, 1950.

Gustafson, D., Shukla, R., Delbecq, A., and Walster, G. A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups, and nominal groups. Organizational Behavior and Human Performance, 1973, 9, 280-291.

Holloman, C., and Hendrick, H. Adequacy of group decisions as a function of the decision making process. Academy of Management Journal, 1972, 15, 175-184.

Jenness, A. The role of discussion in changing opinion regarding a matter of fact. Journal of Abnormal and Social Psychology, 1932, 27, 279-296.

Kahneman, D., and Tversky, A. Subjective probability: a judgment of representativeness. Cognitive Psychology, 1972, 80, 249-260.

Kaplan, A., Skogstad, A., and Girschick, M. The prediction of social and technological events. Public Opinion Quarterly, 1950, 14, 93-110.

Keeney, R. A group preference axiomatization with cardinal utility. Research Memorandum RM-75-47, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1975.

Keeney, R., and Kirkwood, C. Group decision making using cardinal social welfare functions. Management Science, 1975, 22, 430-437.

Keeney, R., and Raiffa, H. Decisions with Multiple Objectives. New York: Wiley, 1976, in press.

Kirkwood, C. Decision analysis incorporating preferences of groups. Technical Report No. 74, Operations Research Center, M.I.T., Cambridge, Ma., 1972.

Klugman, S. Group judgments for familiar and unfamiliar materials. Journal of Genetic Psychology, 1945, 32, 103-110.

Knight, H. A comparison of the reliability of group and individual judgments. Unpublished masters thesis, Columbia University, 1921.

Lichtenstein, S., Fischhoff, B., and Phillips, L. Calibration of probabilities: The state of the art. In Jungermann, H., and de Zeeuw, G., (Eds.), Proceedings of the Fifth Research Conference on Subjective Probability, Utility and Decision Making. 1976, in press.

Linstone, H., and Turoff, M. The Delphi Method: Techniques and Applications. Reading, Ma.: Addison-Wesley, 1975.

Lorge, I., Fox, D., Davitz, J., and Brenner, M. A survey of studies contrasting the quality of group performance and individual performance, 1920-1957. Psychological Bulletin, 1958, 55, 337-372.

Luce, R.D., and Raiffa, H. Games and Decisions. New York: Wiley, 1957.

Maier, N. Assets and liabilities in group problem solving: The need for an integrative function. Psychological Review, 1967, 74, 239-249.

Morris, P. Bayesian expert resolution. PhD dissertation, University Microfilms No. 72-5959, Ann Arbor, Mi., 1971.

Morris, P. Decision analysis expert use. Management Science, 1974, 20, 1233-1241.

Morris, P. Modeling experts. Unpublished manuscript, XEROX Corporation, Palo Alto Research Center, 1975.

Murphy, A., and Winkler, R. Scoring rules in probability assessment and evaluation. Acta Psychologica, 1970, 34, 273-286.

von Neumann, J., and Morgenstern, O. Theory of Games and Economic Behavior (2nd Ed.). Princeton, N.J.: Princeton University Press, 1971.

Norvig, T. Consensus of subjective probabilities: A convergence theorem. Annals of Mathematical Statistics, 1967, 38, 221-225.

Nunnally, J. Psychometric Theory. New York: McGraw-Hill, 1967.

Pattanaik, P. Voting and Collective Choice: Some Aspects of the Theory of Group Decision-Making. Cambridge, England: University Press, 1971.

Pill, J. The Delphi method: Substance, context, a critique and an annotated bibliography. Socio-Economic Planning Sciences, 1971, 5, 57-71.

Raiffa, H. Decision Analysis. Reading, Ma.: Addison-Wesley, 1968.

Rapoport, Am., and Wallsten, T. Individual decision behavior. Annual Review of Psychology, 1972, 23, 131-175.

Rapoport, An. Two-Person Game Theory: The Essential Ideas. Ann Arbor, Mi.: University of Michigan Press, 1966.

Rapoport, An. N-Person Game Theory: Concepts and Applications. Ann Arbor, Mi.: University of Michigan Press, 1970.

Roberts, H. Probabilistic prediction. Journal of the American Statistical Association, 1965, 60, 50-62.

Rowse, G., Gustafson, D., and Ludke, R. Comparison of rules for aggregating subjective likelihood ratios. Organizational Behavior and Human Performance, 1974, 12, 274-285.

Sackman, H. Delphi assessment: Expert opinion, forecasting and group process. Rand Report R-1283-PR, The Rand Corporation, Santa Monica, Ca., 1974.

Savage, L. The Foundations of Statistics. New York: Wiley, 1954.

Seaver, D., von Winterfeldt, D., and Edwards, W. Eliciting subjective probability distributions on continuous variables. Research Report 75-8, Social Science Research Institute, University of Southern California, 1975.

Shaw, M. Group Dynamics. New York: McGraw-Hill, 1971.

Smith, M. Group judgments in the field of personality traits. Journal of Experimental Psychology, 1931, 14, 562-565.

Spetzler, C., and Stael von Holstein, C.-A. Probability encoding in decision analyses. Management Science, 1975, 22, 340-358.

Stael von Holstein, C.-A. Measurement of subjective probability. Acta Psychologica, 1970, 34, 146-159.

Stael von Holstein, C.-A. An experiment in probabilistic weather forecasting. Journal of Applied Meteorology, 1971, 10, 635-645. (a)

Stael von Holstein, C.-A. Two techniques for assessment of subjective probability distributions - An experimental study. Acta Psychologica, 1971, 35, 378-394. (b)

Stael von Holstein, C.-A. Probabilistic forecasting: An experiment related to the stock market. Organizational Behavior and Human Performance, 1972, 8, 139-158.

Stone, M. The opinion pool. Annals of Mathematical Statistics, 1961, 32, 1339-1342.

Stroop, J. Is the judgment of the group better than that of the average member of the group? Journal of Experimental Psychology, 1932, 15, 550-560.

Toda, M. Measurement of subjective probability distributions. Technical Report No. 3, Institute for Research, Division of Mathematical Psychology, Pennsylvania State University, State College, Pa., 1963.

Tversky, A., and Kahneman, D. Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 1973, 5, 207-232.

Tversky, A., and Kahneman, D. Judgment under uncertainty: Heuristics and biases. Science, 1974, 185, 1124-1131.

Van de Ven, A., and Delbecq, A. Nominal versus interacting group processes for committee decision-making effectiveness. Academy of Management Journal, 1971, 14, 203-212.

Van de Ven, A., and Delbecq, A. The effectiveness of nominal, Delphi, and interacting group decision making processes. Academy of Management Journal, 1974, 17, 605-621.

Vinokur, A. Review and theoretical analysis of the effects of group processes upon individual and group decisions involving risk. Psychological Bulletin, 1971, 76, 231-250.

Weaver, W.T. Delphi, a critical review. Research Report RR-7, Educational Policy Research Center, Syracuse University Research Corporation, 1972.

Wilson, R. The theory of syndicates. Econometrica, 1968, 36, 119-132.

Winkler, R. The quantification of judgment: Some methodological suggestions. Journal of the American Statistical Association, 1967, 62, 1105-1120.

Winkler, R. The consensus of subjective probability distributions. Management Science, 1968, 15, 61-75.

Winkler, R. Scoring rules and the evaluation of probability assessors. Journal of the American Statistical Association, 1969, 64, 1073-1078.

Winkler, R. Probabilistic prediction: Some experimental results. Journal of the American Statistical Association, 1971, 66, 675-685.

Winkler, R., and Cummings, L. On the choice of a consensus distribution in Bayesian analysis. Organizational Behavior and Human Performance, 1972, 7, 63-76.

von Winterfeldt, D., and Edwards, W. Flat maxima in linear optimization models. Technical Report 011313-4-T, Engineering Psychology Laboratory, University of Michigan, 1973.

von Winterfeldt, D., and Fischer, G. Multi-attribute utility theory: Models and assessment procedures. In Wendt, D., and Vlek, C. (Eds.), Utility, Probability, and Human Decision Making. Dordrecht-Holland: Reidel, 1975.

Zajonc, R. A note on group judgments and group size. Human Relations, 1962, 15, 177-180.

Zajonc, R. Social facilitation. Science, 1965, 149, 269-274.

Research Distribution List

Department of Defense

Assistant Director (Environment and Life Sciences)
Office of the Deputy Director of Defense Research and Engineering (Research and Advanced Technology)
Attention: Lt. Col. Henry L. Taylor
The Pentagon, Room 3D129
Washington, DC 20301

Office of the Assistant Secretary of Defense (Intelligence)
Attention: CDR Richard Schlaff
The Pentagon, Room 3E279
Washington, DC 20301

Director, Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

Director, Cybernetics Technology Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

Director, Program Management Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209
(two copies)

Administrator, Defense Documentation Center
Attention: DDC-TC
Cameron Station
Alexandria, VA 22314
(12 copies)

Department of the Navy

Office of the Chief of Naval Operations (OP-987)
Attention: Dr. Robert G. Smith
Washington, DC 20350

Director, Engineering Psychology Programs (Code 455)
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217
(three copies)

Assistant Chief for Technology (Code 200)
Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217

Office of Naval Research (Code 230)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Naval Analysis Programs (Code 431)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Operations Research Programs (Code 434)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research (Code 436)
Attention: Dr. Bruce McDonald
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Information Systems Program (Code 437)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research (ONR)
International Programs (Code 1021P)
800 North Quincy Street
Arlington, VA 22217

Director, ONR Branch Office
Attention: Dr. Charles Davis
536 South Clark Street
Chicago, IL 60605

Director, ONR Branch Office
Attention: Dr. J. Lester
495 Summer Street
Boston, MA 02210

Director, ONR Branch Office
Attention: Dr. E. Gloye and Mr. R. Lawson
1030 East Green Street
Pasadena, CA 91106
(two copies)

Dr. M. Bertin
Office of Naval Research
Scientific Liaison Group
American Embassy — Room A-407
APO San Francisco 96503

Director, Naval Research Laboratory
Technical Information Division (Code 2627)
Washington, DC 20375
(six copies)

Director, Naval Research Laboratory (Code 2029)
Washington, DC 20375
(six copies)

Scientific Advisor
Office of the Deputy Chief of Staff
for Research, Development and Studies
Headquarters, U.S. Marine Corps
Arlington Annex, Columbia Pike
Arlington, VA 20380

Headquarters, Naval Material Command (Code 0331)
Attention: Dr. Heber G. Moore
Washington, DC 20360

Headquarters, Naval Material Command (Code 0344)
Attention: Mr. Arnold Rubinstein
Washington, DC 20360

Naval Medical Research and Development Command (Code 44)
Naval Medical Center
Attention: CDR Paul Nelson
Bethesda, MD 20014

Head, Human Factors Division
Naval Electronics Laboratory Center
Attention: Mr. Richard Coburn
San Diego, CA 92152

Dean of Research Administration
Naval Postgraduate School
Monterey, CA 93940

Naval Personnel Research and Development Center
Management Support Department (Code 210)
San Diego, CA 92152

Naval Personnel Research and Development Center (Code 305)
Attention: Dr. Charles Gettys
San Diego, CA 92152

Dr. Fred Muckler
Manned Systems Design, Code 311
Navy Personnel Research and Development Center
San Diego, CA 92152

Human Factors Department (Code N215)
Naval Training Equipment Center
Orlando, FL 32813

Training Analysis and Evaluation Group
Naval Training Equipment Center
(Code N-00T)
Attention: Dr. Alfred F. Smode
Orlando, FL 32813

Department of the Army

Technical Director, U.S. Army Institute for the Behavioral and Social Sciences
Attention: Dr. J.E. Uhlaner
1300 Wilson Boulevard
Arlington, VA 22209

Director, Individual Training and Performance Research Laboratory
U.S. Army Institute for the Behavioral and
and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

Director, Organization and Systems Research Laboratory
U.S. Army Institute for the Behavioral and
Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

Department of the Air Force

Air Force Office of Scientific Research
Life Sciences Directorate
Building 410, Bolling AFB
Washington, DC 20332

Robert G. Gough, Major, USAF
Associate Professor
Department of Economics, Geography and
Management
USAF Academy, CO 80840

Chief, Systems Effectiveness Branch
Human Engineering Division
Attention: Dr. Donald A. Topmiller
Wright-Patterson AFB, OH 45433

Aerospace Medical Division (Code RDH)
Attention: Lt. Col. John Courtright
Brooks AFB, TX 78235

Other Institutions

The Johns Hopkins University
Department of Psychology
Attention: Dr. Alphonse Chapanis
Charles and 34th Streets
Baltimore, MD 21218

Institute for Defense Analyses
Attention: Dr. Jesse Orlansky
400 Army Navy Drive
Arlington, VA 22202

Director, Social Science Research Institute
University of Southern California
Attention: Dr. Ward Edwards
Los Angeles, CA 90007

Perceptronics, Incorporated
Attention: Dr. Amos Freedy
6271 Variel Avenue
Woodland Hills, CA 91364

Director, Human Factors Wing
Defense and Civil Institute of
Environmental Medicine
P.O. Box 2000
Downsville, Toronto
Ontario, Canada

Stanford University
Attention: Dr. R.A. Howard
Stanford, CA 94305

Montgomery College
Department of Psychology
Attention: Dr. Victor Fields
Rockville, MD 20850

General Research Corporation
Attention: Mr. George Pugh
7655 Old Springhouse Road
McLean, VA 22101

Oceanautics, Incorporated
Attention: Dr. W.S. Vaughan
3308 Dodge Park Road
Landover, MD 20785

Director, Applied Psychology Unit
Medical Research Council
Attention: Dr. A.D. Baddeley
15 Chaucer Road
Cambridge, CB 2EF
England

Department of Psychology
Catholic University
Attention: Dr. Bruce M. Ross
Washington, DC 20017

Stanford Research Institute
Decision Analysis Group
Attention: Dr. Allan C. Miller III
Menlo Park, CA 94025

Human Factors Research, Incorporated
Santa Barbara Research Park
Attention: Dr. Robert R. Mackie
6780 Cortona Drive
Goleta, CA 93017

University of Washington
Department of Psychology
Attention: Dr. Lee Roy Beach
Seattle, WA 98195

Eclectech Associates, Incorporated
Post Office Box 179
Attention: Mr. Alan J. Pesch
North Stonington, CT 06359

Hebrew University
Department of Psychology
Attention: Dr. Amos Tversky
Jerusalem, Israel

Dr. T. Owen Jacobs
Post Office Box 3122
Ft. Leavenworth, KS 66027

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER [redacted]	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Assessment of Group Preferences and Group Uncertainty for Decision Making		5. TYPE OF REPORT & PERIOD COVERED Technical 7/75 - 9/76 6. PERFORMING ORG. REPORT NUMBER SSRI 76-4
7. AUTHOR(s) David A. Seaver		8. CONTRACT OR GRANT NUMBER(s) Prime Contract #N00014-76-C0074 Subcontract #75-030-0711
9. PERFORMING ORGANIZATION NAME AND ADDRESS Social Science Research Institute University of Southern California Los Angeles, CA 90007		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Advanced Research Projects Agency 1400 Wilson Boulevard Arlington, Virginia 22209		12. REPORT DATE June, 1976 13. NUMBER OF PAGES 66
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Office of Naval Research 800 North Quincy Street Arlington VA 22217		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approval for public release, distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Support for this research performed by Social Science Research Institute was provided by the Advanced Research Projects Agency of the Department of Defense and was monitored under Contract N00014-76-C-0074 with the Office of Naval Research, under subcontract from Decisions and Designs, Inc.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

decision analysis	Pareto optimality	statisticized group
group decision making	group judgment	
subjective probability	Delphi method	
expected utility	nominal group	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Often groups rather than individuals function as decision makers. In order to use decision analysis to aid such decision makers, the group preferences and opinions must be quantified as utilities and probabilities. This paper reviews the procedures by which these group utilities and probabilities can be determined, along with the relative merits of individual versus group judgments of these quantities. Research has shown quite conclusively that group judgments are generally superior to individual judgments, but no entirely

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

satisfactory method for determining either group utilities or probabilities exists. Using ordinal preferences rather than cardinal utilities, majority rule will lead to a satisfactory group preference function in certain restricted situations. Other possibilities include use of anchored scales and weighted averages of individual cardinal utilities. Both mathematical aggregation rules and behavioral approaches have been suggested for obtaining group probabilities. The mathematical approaches range from the simplicity of weighted averaging to much more complex Bayesian methods. The theoretically elegant methods, however, generally suffer in practice from the impossibility of determining some of the inputs necessary for the models. The behavioral approaches attempt to reduce disagreement by communication and interaction among the group members. The most successful methods depend on structured communication to allow the facilitation of group judgments while avoiding many of the detrimental influences that have been identified in social psychological research.